THE UNIVERSE OF NEUROTOXIC PROTEINS:
A STUDY OF THE CONFORMATIONAL SPACE OF POLYGLUTAMINE AND B-AMYLOID

Àngel Gómez-Sicilia

Supervisor: Dr. Mariano S. Carrión Vázquez (Cajal Institute – CSIC)
Supervisor: Prof. Marek Cieplak (Institute of Physics – PAN)

Tutor: Prof. Julio Gómez Herrero (Dept. of Condensed Matter Physics – UAM)

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Biophysics in the Department of Condensed Matter Physics.

July 2015.

Acknowledgements

Over the journey that this thesis has supposed many people have crossed paths with me who need to be present in this work, for without them none of it would have been possible.

Firstly, I address my thanks to Mariano. It is not only because you provided me with the tools that I needed in the development of this work, but also that through fruitful discussions and sometimes heated arguments we were able to generate several works together. Thank you for carrying me from Physics to Biology and presenting me with a world of questions ready to be answered.

Next in the list is Marek. You I thank for bringing me back to Physics and theory, for your patience when my priorities have had to be somewhere else and for giving me the chance to discover a new country, a different working style and opera.

This work could not have been done without the great contribution of the supercomputing clusters at CSIC Trueno, Teide High Performance Computing and Galician Supercomputing Center, particularly Mr. Arurelio Rodríguez, who had a major contribution to software development. Furthermore, I must thank the funding entities responsible for my maintenance: CiberNed, CSIC through a JAE-Pre scholarship and the project 7FP-NMP ref. 604530.

Next I would like to thank the former members of the lab. Firstly Javier, Álex and Oscar G, who left long before I finished, but who taught me both the techniques and the abilities I needed to reach this point. Also those that went through the lab even if it was only for a short time, each of which left a little mark in it: Fernando, Amalia, Débora, Elena, Aida, Lara, Clara, Bárbara, Miguel, Natalia, Francisco, Laura, María, Elena and Oscar S.

There are also the people who I currently share the lab with, starting with Rubén, thank you for your songs, your teasing, the moments where you went serious and the Japanese word “som-hi”; Andrés, thank you for your experimental meticulousness, your disordered table and learning to speak Catalan by saying “això”; Albert, thank you for sharing the journey nearly from the beginning, for making me run before meals and after, and for teaching me how to prepare ubiquitin that behaves like a cohesin in the AFM; Mari Carmen, thank you for being a great partner, nagging at the right spots but mostly patient, and a special thanks for being my style guide, probably the best I can ever get; Isabel, thank you for acting as my mother one second and partying the night off the next, for being so straight and standing up for all of us; Laura, thank you for the management

work, which needs to be explicitly recognized, but also for making me recall my university days with long discussions about math, languages and tools; Emilio, thank you for surprising me with your thousand voices as well as your culinary taste and for the never-ending discussions on whatever topic comes up; Adrián, thank you for being so innocent, I had forgot that part of my life and I really miss it; Juan Carlos, thank you for your endless energy, which was contagious enough to activate the rest of us; and Roberto, thank you for the enthusiasm you show with your daily work, which greatly increases my motivation. Also the people at IFPAN, Michał, Karol, Bartosz, Adolfo and Mateusz; thank you for the warm welcome, even if it was just for a few months, and for all the help I had from you during that time and after.

Out of the lab, former and current "concienzudos" have a lot of responsibility in this thesis as well: Natalia, Fran, Ana, Andrea, Julián, Adri A, Adrián B, Ángel T, Laura, Merce, Jaime and Míriam. Furthermore, I also thank the people at the Cajal Institute services and management who make it run as smoothly as possible. Finally, this thesis has received insightful feedback from Douglas and Miguel at the Rocasolano Institute, whom I further warmly thank.

I cannot close this chapter without heartfeltly thanking José. Thank you for putting up with me in the days I'm off. Thank you for forcefully taking me out of the box to give me new sights of the things I am staring at. Thank you for taking me out to Friday's on Thursdays even if it's probably not your first choice and for making me discover C, which led me to develop a great interest in programming on which this work is grounded. For this and much more, thank you.

Last but not least, my gratitude goes to my friends and family. With some I shared The Beach: Laia, Berta, Cris, Sergi, Juanjo, Cristina, Esther, Irene, Lorena… With others I split my studies: Saül, Anna, Sònia, Isa, Ramon, Miquel, Gugli, Pol… From some I received welcome: Irene, Lara, Rosa, Elva, Lina, Marta, Ángel, Sergio, Aitor, Àlex, Sílvia, Richi… And with most I shared and hope to keep sharing alcohol. Nonetheless, I experienced the whole journey – or a great deal of it – with my family, especially my brothers Carlos and Dani, and my parents, who deserve to know how much of their help and support was vital for the completion of this work.

TO MY FIVE GRANDPARENTS,
One that showed me the importance of the unknown,
One that taught me the value of family,
One that gave me the power of thought,
One that guided me into belief, and
One that nurtured me in the art of dining.

Summary

Neurodegenerative diseases are considered diseases of the old. These are among the main causes of death after the discovery of vaccination and antibiotics in the late 18th and early 20th century, respectively, pushed the average life expectancy in the developed countries to the current age of about 80. Among them, Alzheimer and Huntington are incurable diseases, even though a great effort has been put in their understanding and several important discoveries have been made about their causal factors.

Both diseases are related to specific proteins in the nervous system: β-amyloid (Alzheimer) and polyglutamine expansions in huntingtin (Huntington). Polyglutamine expansions are also present in other proteins, which are in turn involved in several other diseases such as Spinocerebellar Ataxias. The study of these rapidly fluctuating proteins is challenging using the current biophysical techniques, which do not give access to the whole distribution of conformations but average over the whole population of molecules and thereby hide rare events. The advent of single-molecule techniques, as well as their good correlation with computer simulations, has made the atomistic exploration of these proteins possible.

In this work, we combine atomic force microscopy with molecular dynamics simulations to study the conformational polymorphism of polyglutamine expansions and β-amyloid. We discover that both proteins have very similar behaviour at the monomeric level even if their amino acid sequences are quite different. When allowed to evolve, these proteins continuously waver between several different conformers that have distinct shapes and properties. Among these conformations, we find some that last longer than others, some that present a high resistance to forced unfolding and, in simulations, some that generate a knot in their structure. With these findings, we propose that high temporal stability, high stability under force and the presence of knots might explain the toxicity of these proteins at the monomeric level, since the proteasomal degradation of some of these species seems to be troublesome and, under certain circumstances, impossible.

Resumen

Las enfermedades neurodegenerativas se consideradan enfermedades de la vejez. Éstas se encuentran entre las mayores causas de mortalidad desde que el descubrimiento de las vacunas y los antibióticos a finales del siglo XVIII y a principios del XX, respectivamente, alargaron la esperanza de vida (en paises desarrollados) hasta la edad actual de alrededor de 80 años. Entre ellas, el Alzheimer y el Huntington son enfermedades incurables, a pesar del gran esfuerzo realizado para entenderlas y los distintos descubrimientos realizados acerca de sus factores causales.

Ambas enfermedades están relacionadas con proteínas del sistema nervioso: β-amiloide (Alzheimer) y expansiones de poliglutaminas en la huntingtina (Huntington). Las expansiones de poliglutaminas también están presentes en otras proteínas, que a su vez están implicadas en otras enfermedades como las Ataxias Espinocerebelares. El estudio de estas proteínas rápidamente fluctuantes presenta un reto para las técnicas actuales de biofísica, que no dan acceso a la distribución completa de posibilidades sino que promedian sobre la población y con ello esconden eventos raros. El advenimiento de las técnicas de molécula individual, junto con su buena correlación con las simulaciones por ordenador, ha hecho posible la exploración de estas proteínas.

En este trabajo combinamos la microscopía de fuerzas atómicas con simulaciones de dinámica molecular para estudiar el polimorfismo conformational de las expansiones de poliglutamina y el β-amiloide. Descubrimos que el comportamiento de ambas es similar a nivel de monómero a pesar de que sus secuencias aminoacídicas son distintas. Cuando se las deja evolucionar, estas proteínas fluctúan rápidamente entre varios confórmeros que presentan formas y propiedades distintas. Entre estas conformaciones, encontramos algunas que duran más que otras, varias con alta resistencia al desplegamiento bajo fuerza y, en simulaciones, algunas que generan un nudo en su estructura. Con estos descubrimientos, proponemos que la estabilidad temporal, la alta estabilidad bajo fuerza y la presencia de nudos podrían explicar la toxicidad de éstas proteínas a nivel de monómero, ya que la degradación de algunas de estas especies en el proteasoma parece ser problemática y, en ciertas condiciones, imposible.

Contents

List of Figures

List of Tables

List of abbreviations

1BNR	Barnase
1C4P	β domain of streptokinase
1G1K	Cohesin module from *Clostridium cellulolyticum*
1GB1	Inmunoglobulin-binding domain of streptococcal protein G
1TIT	I27 domain of titin
1UBQ	Ubiquitin
Aβ	β-amyloid
AFM	Atomic Force Microscopy, Atomic Force Microscope
ATP	Adenosine triphosphate
BEMD	Bias Exchange Molecular Dynamics
C_α	Alpha carbon – The backbone carbon atom that links to the side chain
C_β	Beta carbon – The carbon atom in the side chain that is bound to the C_α
CDF	Cumulative Density Function
CSU	Contact for Structural Units
DNA	Deoxyribonucleic Acid

FC	Force-Control
$F_{\mathbf{max}}$	mechanical stability
G	Guest
H	Host
I27	the 27th inmunoglobulin module from human cardiac titin (original nomeclature)
IDP	Intrinsically Disordered Protein
LC	Length-Control
M	single-molecule Marker
MD	Molecular Dynamics
oCSU	Original CSU
OV+oCSU	OV combined with original CSU
OV	overlap
PDB	Protein Data Bank
PDF	Probability Density Function
QBP1	Glutamine Binding Peptide 1
$\mathbf{Q}_n$	polyglutamine
rCSU	Repulsion CSU
OV+rCSU	OV combined with repulsion CSU
RMSD	Root of the Mean Square Deviation
RNA	Ribonucleic Acid
SMFS	Single Molecule Force Spectroscopy
$\mathbb{SS}$	Secondary Structure Content
WLC	Worm-Like Chain

Part I

Preface

1. Introduction

Renaissance scientists were people with many fields of expertise, including life sciences, astronomy and mathematics. Later in human evolution, especially after the scientific revolution in the 19th century, many scientific discoveries were made which led to a better and more profound understanding of the world and the creatures that live in it, but also forced anyone interested in majoring on a field to specialize in order to reach that profound understanding. Thus, science as a whole diverged into several fields such as physics, chemistry or biology, each of which branched in turn into smaller areas as different as astronomy, material science or microbiology.

Nonetheless, a deeper knowledge of the world led only to the rise of more complex questions, especially ones that might belong in the frontier of these branches. One could think of a simple example: Proteins, responsible for living and present in living organisms (and thus typically studied by biologists) fold into a structure based mainly on the chemical properties of the side chains in their amino acid residues (and thus should belong to chemistry department) to carry out functions such as transport, anchoring or force transmission (all of them related to mechanics, that is, physics). In the spirit of understanding these new questions, each separate individual needs to think further away from their own

field of expertise and use knowledge that comes from other fields, be it as a tool for further digging or as a shift in the point of view. It is with this collaborative spirit that this thesis is on biophysics.

1.1. The folding landscape of proteins

One of the oldest questions in molecular biology, dating as back as mid twentieth century, is to find out how macromolecules of living organisms (namely DNA, RNA and proteins) self-arrange into organized packaged three-dimensional structures, or folds. The problem was solved for DNA [1], with a four-base composition, resulting into a double helical structure of complementary branches; yet even if somehow similar structures were proposed at the time for proteins [2], the problem of protein folding is still a challenge which has brought into biology many techniques from other fields such as physics or chemistry.

The importance of this problem comes from the thought, derived from the central dogma of molecular biology (see Fig. 1.1), that each protein has one or more functions associated with it which derive from its fold. With this idea in mind, being able to understand the relationship between folds and functions, and being able to recreate these folds are two things that are highly desirable to accomplish. However, since proteins are typically linear chains composed of amino acid residues of 20 different kinds, the study of the energetics responsible for this matter is, as of today, still not possible.

The folding of a protein is known to depend on many factors other than the sequence of the protein itself. Indeed, environmental conditions such as temperature, pH, solvent and others can act on a fully folded protein and unfold it, or *vice versa*. Ribonuclease A was the first molecule to be unfolded and refolded *in vitro* [3], signaling the start of the modern folding theories: It was Anfinsen who postulated that the folding state of a protein needs to be the one in which the Gibbs free energy of the system is minimal [4]. This theory, thus, proposed that a protein and its surrounding could (and eventually would) find one and only one minimum in an otherwise flat energy landscape.

About the same time [5], Levinthal came up with the following thought experiment: Let us assume a protein with 100 residues, each of which can adopt three different conformations. Let us further assume that the time needed to explore one conformation is on the order of tenths of picoseconds ($\simeq 10^{-13}$ s). The timescales that such a system would need to explore all possible conformations

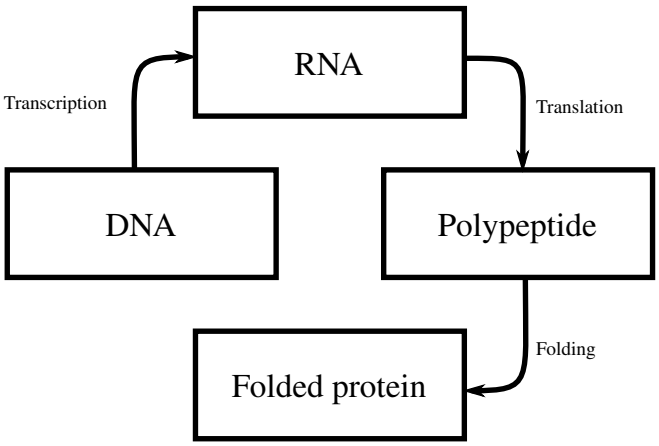


Figure 1.1: Central dogma of molecular biology. DNA encodes information that is read by the RNA polymerase, which transcribes it into messenger RNA. This is in turn read by the ribosome, which translates it into a polypeptidic chain. Lastly, the chain (either on its own or with the help of some molecular chaperones) undergoes a folding process to become a folded protein, which carries out a function.

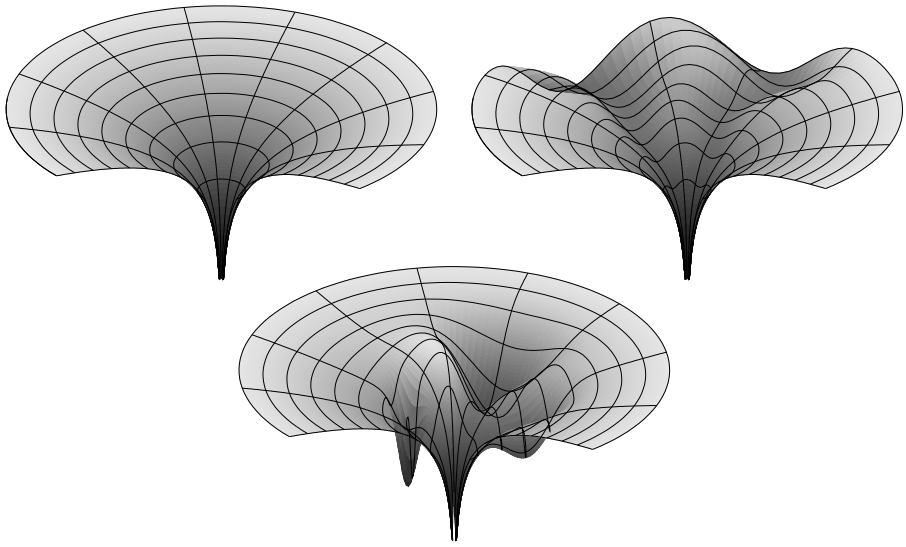


Figure 1.2: Energy landscape models. The simple funnel model (left) [4] states that a protein has only one folded state and will eventually find it. Levinthal's paradox (right) [5] involves finding preferred paths where the protein is easily directed to the minimum. Typical energy landscapes (bottom) present several local minima known as folding intermediates, which can be on- or off-pathway, the latter typically leading to misfolding [6].

is $3^{100} \cdot 10^{-13}$ s $\simeq 5 \cdot 10^{34}$ s, 10^{17} times the age of the universe. However, typical globular proteins of such a size undergo their folding process in timescales of minutes at most (some fast-folder proteins fold in microsecond timescales). Thus, evolution has come up with a way of resolving this so-called Levinthal's paradox, which implies the presence of preferred exploratory pathways for unfolded proteins to reach their folded state.

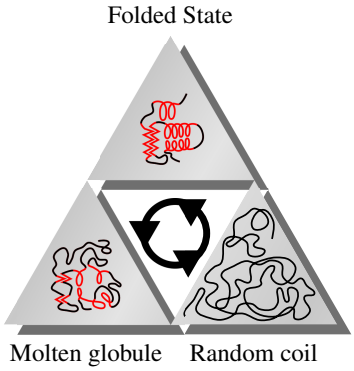
Moreover, contrary to the original belief that the folding of a protein is a two-state equilibrium process where only the folded and unfolded species existed, many kinetic experiments designed to solve Levinthal's paradox prove the existence of other species, called folding intermediates, which correspond to several minima in the Gibbs free energy. These metastable states can be located either in or out of the preferred folding route (or routes). The former, typically called on-pathway, represent states to which the unfolded molecule easily travels, and from which it easily leaves to reach lower-energy states (which can be other intermediates or the folded state). On the other hand, minima located out of the folding route are often referred to as off-pathway states and their study gives information regarding trap states to which the unfolded protein can arrive but from which it cannot easily leave, thus never reaching a fold and generating a misfolded structure [6].

In order to explain Levinthal's paradox, several models have been proposed for different stages of the folding process, including the nucleation-propagation model [7], the framework model [8], the diffusion-collision model [9] and the hydrophobic collapse model [10].

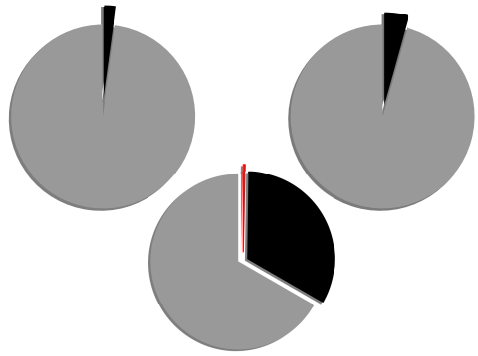
1.2. Intrinsically Disordered Proteins

Currently, the central dogma (see Fig. 1.1) is known to apply only to a fraction of the known proteins. Indeed, there is a group of proteins that do not fold immediately after translation, but present an equilibrium in which they sample the conformational space acquiring many transient structures and acquire a fold only upon binding to a partner, if at all [11]. These proteins are typically known as Intrinsically Disordered Proteins (IDPs) [12], and they comprise proteins that are completely disordered as well as those that have long regions in their sequence with no specific fold – more than 30 residues.

Even if the name suggests that IDPs are always in disordered states, they are known to explore an energy landscape where they can acquire secondary



(a) Protein trinity



(b) IDPs in the proteome

Figure 1.3: Intrinsically Disordered Proteins (IDPs). (a) The protein trinity is a new paradigm that establishes that proteins fluctuate between three states: The folded state, the molten globule and the random coil. Secondary structure elements are highlighted in red. (b) Fraction of the archaea (left), eubacteria (right) and eukaryotic (bottom) proteome formed by IDPs. In black, the fraction occupied by IDPs; in gray, the rest of the proteins. The red section in the eukaryotic graph represents the fraction of IDPs involved in conformational diseases.

structure elements temporarily. These secondary structure elements are vital in the interaction with their ligands [13]. However, the average behaviour of these proteins is consistent with unfolded polymers.

After the discovery of IDPs, a new paradigm complements the central dogma, whereby the proteins live in a constant equilibrium going from an ordered state to a molten globule state – where the protein has the size of the ordered protein, but the secondary structure elements are not yet formed – and from there to a random-coiled state – where the molecule can extend and explore the space in order to fold again. This new paradigm, represented in Fig. 1.3a, is known as the protein trinity [14].

With the computing power available today, predictions have been made as to what fraction of the proteome would be formed by IDPs (see Fig. 1.3b). Interestingly, the eukaryotic proteome presents 33.0 % of its proteins completely or partially disordered. This fraction is different than that of the proteins from archaea and eubacteria, that have respectively 2.0 % and 4.2 % of their proteome formed by IDPs [15]. This fraction of the proteome complements the function of

the ordered proteins and is critical in several biological functions such as cellular signalling.

These proteins are a great evolutionary advantage, since they can change conformation rapidly and thus present a high binding surface for such small proteins. This is why most of them are involved in one or more biological functions. Nonetheless, IDPs are much more involved in disease than folded proteins. In particular, they are involved in many conformational disorders, as well as non-conformational ones such as cancer or diabetes. Among the conformational diseases, neurodegenerative diseases are prominent. These diseases are caused by a toxic gain of function in specific proteins, known as neurotoxic proteins, that cause neuronal death. Neurotoxic proteins, besides the toxic gain of function, are involved in amyloidogenesis – a process by which they bind to one another, firstly generating soluble oligomers that then continue aggregating to spawn insoluble fibres [16, 17].

Representative neurotoxic proteins include α-synuclein (related to Parkinson), β-amyloid (linked to Alzheimer), polyglutamine expansions (associated to Huntington among other diseases) and tau (involved in many disorders including Alzheimer and Huntington). The fact that the neurotoxic and amyloidogenic properties of these proteins are related to disease leads to think that a better understanding of them at the molecular level will shed light on the mechanisms involved with the diseases [17]. In this thesis, we try to discover new insights in two of these proteins: polyglutamine (Q_n, where n stands for the number of glutamines in the chain) and β-amyloid (Aβ).

1.2.1. Polyglutamine and Huntington disease

Q_n is one of the many homopolymeric tracts present in the eukaryotic proteome. In particular, 18.9 % of the human proteome involves homopolymeric tracts of size 5 or greater, while the probability of one happening by chance is $6 \cdot 10^{-6}$ (data obtained from Ref. [18]). Nonetheless, Q_n constitute the second longest such chains, with FoxP2, a protein related to human language, having 40 repeats (random probability of $9 \cdot 10^{-53}$).

Q_n chains can be found in many proteins. One example is huntingtin, a protein known to be involved in development [19], and thought to be related to gene expression regulation [20] and to anchoring or transport of vesicles [21], although its function is not completely elucidated. Nonetheless, this protein is also known to undergo a toxic gain of function and be directly related to Huntington

disease.

Huntington disease, also called Huntington corea (from Greek χορεία, dance), affects one every 20 000 people, onsets between the ages of 30 and 45 and patients die 10 to 20 years thereafter. Its symptoms mainly encompass coreic – chaotic and involuntary – movements, followed by aggressivity and dementia including depression, lack of concentration and loss of short-term memory [22].

Other Q_n-containing proteins are known to be involved in several other diseases, collectively known as polyglutaminopathies, such as some Spinocerebellar Ataxias, Dentatorubropallidolusyan Atrophy, and Spinal and Bulbar Muscular Atrophy. All these proteins, at the DNA level, are known to undergo trinucleotide-repeat expansion, a DNA mutation caused by slippage of the DNA polymerase [23]. The mutation results in abnormally long repetitions of the same three bases which, after translation, generate a mutant protein with an expansion of Q_n longer than the wild-type form. In some cases, the number of glutamines will exceed a certain (disease-dependent) threshold and will become toxic to the cells and lead to disease. The thresholds for Q_n diseases have a median of 35, which is also Huntington disease threshold. Interestingly, although this threshold is much smaller in the case of Spinocerebellar Ataxia 7 ($n = 17$) and much larger for Spinocerebellar ataxia 17 ($n = 42$), most of the other thresholds are close to $n = 35$. This fact marks an important difference with other similar diseases such as Alzheimer: polyglutaminopathies are not sporadic but genetically determined.

A great share of experimental effort has been made to study the negative effects of neurodegenerative diseases at multiple levels, from *in vivo* studies [24] to *in vitro* ones focusing on the fibres [25] or the oligomers [26] that are formed in the amyloidogenic pathway. However, these studies have not yet been able to elucidate the particular toxic species or, in the case of homopolymeric tracts, the origin of the threshold in the length of the disease-inducing ones.

1.2.2. β-amyloid and Alzheimer disease

Another disease-implicated IDP is Aβ. Aβ is a peptide derived from the processing of the amyloid precursor protein, located at the plasmatic membrane [27]. It is one of the proteins found in the autopsy of Alzheimer disease patients, in the form of extracellular aggregates, along with intracellular tau aggregates.

Alzheimer disease is characterized by a progressive decline in cognitive function, specifically degrading remembering, learning, reasoning and communication capabilities [28]. Its prevalence is around 1 % for patients below the age of

65, but increases exponentially with the age and reaches 30 % at age 85 [29]. In addition to sporadic Alzheimer, mutations have been discovered in the genes encoding for amyloid precursor protein which induce greater severity or earlier onset of Alzheimer disease in patients [30].

We center this work on the study of Aβ, which is considered to be the main determinant in Alzheimer. Aβ peptide is, as aforementioned, cut off of amyloid precursor protein, but the final length is not fixed, ranging from 38 to 42 residues. The most abundant species are $A\beta_{40}$ and $A\beta_{42}$, the former being much more abundantly secreted by healthy neurons. The study of the wild-type Aβ of lengths 40 –non-toxic– and 42 –toxic– is complemented in this work with studies of mutants with several characteristics. These mutants[1] are E3R $A\beta_{40}$, F19S/L34P $A\beta_{42}$, E22G $A\beta_{42}$ and E22G/I31E $A\beta_{42}$. These are known to induce toxicity, prevent aggregation, be more aggressive in Alzheimer patients and promote fast fibrillation avoiding toxicity, respectively [31, 32].

1.2.3. The study of IDPs: Single Molecule techniques

Classical biochemical and biophysical techniques deal with large populations of proteins at once. This is the case of nuclear magnetic resonance spectroscopy, circular dicroism spectroscopy and X-ray crystallography, among others. These techniques have been used extensively to study IDPs and have yielded information on aggregation, fibrillogenesis and the structure of mature fibers [25, 33], but the fast fluctuations together with the heterogeneity in the sample have made the monomer elusive.

Indeed, bulk techniques are known to sample an average of the several conformations in a system – averaging typically as many as Avogadro's number. Such averages, as stated by the central limit theorem, are normally distributed independent of the probability distribution of each sample, and therefore cases where the system resides longer are favoured to the detriment of the less frequent states [34].

Single molecule techniques can be used to solve this issue. Using them we can measure the properties of one molecule at a time, and therefore study the whole population and obtain a probability density of the sample instead of the average. Thus, we can obtain information of several possible states of the molecule,

[1] Mutations are noted as follows: X*n*Z means residue X at position *n* becomes residue Z, where X and Z are the amino acid 1-letter codes.

as well as many different pathways for reactions such as protein unfolding. In particular, single molecule fluorescence and Atomic Force Microscopy (AFM, this abbreviation will refer both to the technique and the device) have been used to characterize IDPs in terms of the study of their conformational dynamics and conformational polymorphism [35].

1.3. Mechanics in biology

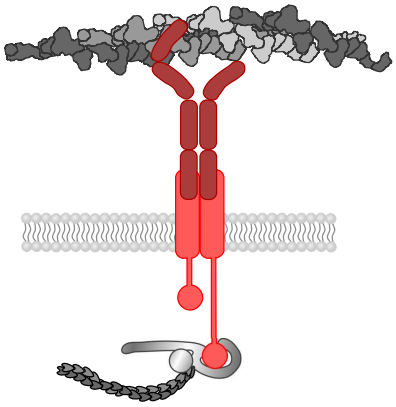
The first question that rises in biophysics is what kind of relationship can physics and biology establish in order to further push the frontiers of science. In order to answer this question, we simply need to look at the cell, unit of a living system, as a factory where nanometer-scale devices work in sync to achieve complex tasks with the common goal of life. These devices are mainly proteins, linear polymers composed of twenty different types of amino acid residues, which interact with one another typically in an ordered fashion to give rise to a well-defined three-dimensional structure known as the protein fold. Some proteins are involved in transforming chemical energy obtained from ATP hydrolysis into mechanical energy in order to carry out functions such as adhesion, transport or movement, among others. Therefore, studying the mechanics of the living systems is crucial to be able to understand them.

Some proteins are mainly involved in an obviously mechanical function, and appear to have evolved in order to generate or sustain force. Here are some examples (see Fig. 1.4): titin is a giant protein formed by several modules that is present in the muscle cells and is responsible for their passive elasticity [36]; kinesin is a motor protein that is in charge of dragging vesicles from the nucleus of the cell to its periphery in eukaryotes along microtubules [37]; fibronectin is present in the extracellular matrix and binds to integrins on the cell membrane in order to facilitate migration and cell adhesion [38].

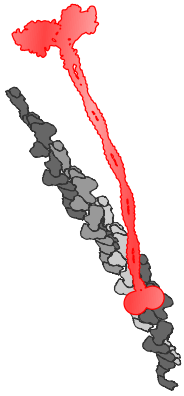
There are many other proteins which are not so clearly involved in mechanical functions, such as DNA polymerase, which replicates DNA by gathering nucleotides [39]. Nonetheless, most of the proteins in eukaryotic cells, archaea and bacteria are unfolded by pulling at some point, either due to cross-membrane translocation through a pore [40] or to degradation through the ubiquitin-proteasome pathway [41].

Therefore, knowing the forces involved with each protein is of vital importance for the understanding of their regulation.

(a) Titin



(b) Fibronectin – Integrins



(c) Kinesin

Figure 1.4: Mechanics in biology. (a) Titin –in red– is a giant protein of the muscle cells, where it plays the role of a biological spring in recovering the original shape after stretching. (b) Fibronectin –in dark red– and integrins –in light red– connect cells to the extracellular matrix, and they contract and expand helping to accomplish functions such as cell adhesion and movement. (c) Kinesin –in red– is a motor protein that transports vesicles by dragging them along microtubules.

1.3.1. Protein unfolding machinery

All the motor machines involved in the translocation are known to function by mechanically pulling from one of the ends of a protein against a narrow pore which does not allow for the entering of a globular protein maintaining its three-dimensional structure and thus forcefully unfolds the folded protein [42].

In particular, protein degradation in the cell is key in removing damaged proteins, preserving enzymatic activity and controlling transcription factors, among other processes [43]. This process is typically carried out in multisubunit proteases, multiprotein complexes performing a two-step process of first unfolding the protein and then cutting the peptide into small pieces to be reutilised later.

In archaea and eukaryotic cells, this proteases are known as proteasomes, the most common being the 26S, composed of the 19S unfolding unit and the 20S digesting unit [44]. Nonetheless, for the sake of simplicity, the typical model chosen for the study of protein degradation is ClpXP from bacteria, similarly composed by the ClpX unfoldase ring and the ClpP proteolytic chamber, made only by two types of subunits [45, 46].

The unfolding unit of most proteases is actually a complex itself, composed by six identical subunits forming a hexamer. This unit actively generates a pulling stroke thanks to the consumption of energy from ATP-hydrolysis, which induces a conformational change. This stroke not only pulls towards the inside of the chamber, but also rotates the subunit subjecting the protein to torsional forces that might further help in the unfolding [47].

Part II

Techniques

2. Atomic Force Microscopy

The experimental part of this work was carried out using an AFM, mainly in its force spectroscopy mode. AFM is a technique suited to study the mechanical stability of proteins in the high-forces regime (typically 10 to 1000 pN) with a very high precision in distance measurements (down to 0.1 nm).

2.1. Origins and history

The AFM belongs to the family of techniques known as scanning probe microscopy. The first of these techniques ever applied to a biological sample was scanning tunneling microscopy [48], which is based on the transmission of a current from a tip into a conducting substrate hinged on the tunnel effect. The experiment in question was acquiring high resolution images of some purified viral particles and oligomers from the collar of the bacteriophage Φ29. Nonetheless, scanning tunneling microscopy came with several handicaps in terms of its use on biological samples, namely that the sample and substrate need to be conducting to allow the tunnel effect, that images need to be taken in vacuum or air since physiological buffer is conducting and would not allow the tunnel effect to happen, and that the sample is directly exposed to a current, however low, which

might have an effect on its properties.

Many techniques were invented to overcome these limitations, such as Kelvin probe force microscopy, magnetic resonance force microscopy or scanning thermal microscopy, to name a few. However, to date, the most used techniques in this family are scanning tunneling microscopy and AFM.

AFM was originally invented as an imaging tool, but had the advantage of measuring biological samples in their physiological buffer without the need to apply any current. The first sample imaged with AFM was a ceramic made of aluminium oxide [49], but it rapidly shifted to biological samples all the way from cell morphology to protein structure [50].

After several years of its invention for imaging, AFM was applied for the first time to force spectroscopy. In particular, the first experiment [51] was done on human cardiac titin, a giant protein that is responsible for the elasticity of the sarcomere in muscle cells, which is composed of several independently-folded modules. This protein was chosen because of the *pseudo*-periodicity that it presents, and it was pulled with constant speed, producing force-extension plots similar to a saw. This kind of patterns is nowadays standard in the field of AFM-based Single Molecule Force Spectroscopy (SMFS), and was christened *sawtooth pattern*. Interestingly, the same experiment was carried out simultaneously using optical tweezers, another force spectroscopy technique, on the same protein yielding similar results [52].

2.2. The set-up

The AFM set-up consists of two parts. Firstly, there is a force sensor consisting of a deflectable cantilever on which a laser beam is reflected onto a photoelectric detector. Fig. 2.1a shows a close-up image of the cantilevers and a cartoon explaining the detection mechanism. The photoelectric detector is split in four sectors, namely a, b, c, d, from left to right then top to bottom, and is capable of measuring light intensity in each section as the photoelectric effect potential. The total intensity of the laser, V_Σ, is computed as the sum of the intensities received on each section (Eq. 2.1). The cantilever deflection (related to the normal force, V_N) is computed as the difference between the two top sections and the two bottom ones (Eq. 2.2), while its torsion (related to the lateral force, V_L) is measured as the difference between the two left sections and the two on the right (Eq. 2.3). These two are normalized to the total laser intensity, so that fluctuations in the

laser intensity during an experiment as well as changes in reflectivity between cantilevers in different experiments do not affect the measurement significantly. The reflection of the laser beam on the cantilever, as well as the position of the fotodetector, can be slightly adjusted so that at rest the beam reflects on the center of the detector and therefore both the normal and the lateral forces are zero.

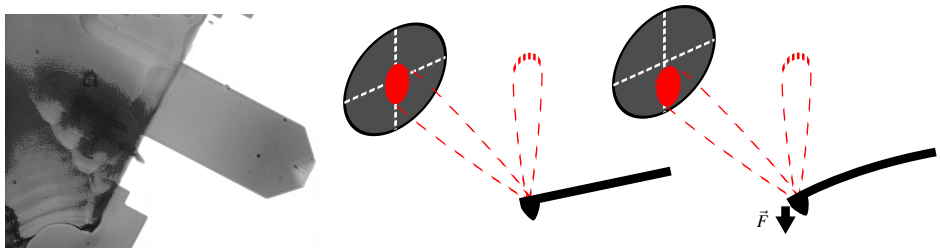
$$V_\Sigma = V_a + V_b + V_c + V_d \tag{2.1}$$

$$V_N = \frac{(V_a + V_b) - (V_c + V_d)}{V_\Sigma} \tag{2.2}$$

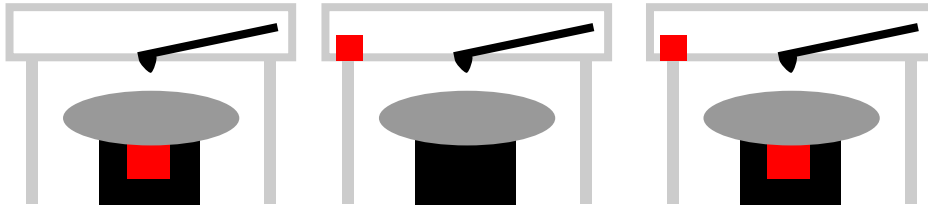
$$V_L = \frac{(V_a + V_c) - (V_b + V_d)}{V_\Sigma} \tag{2.3}$$

The other important part in the set-up is a precise piezoelectric actuator, which is in charge of approaching or distancing the sample relative to the aforementioned cantilever sensor. This piezoelectric positioner can be placed either to move the sample or to move the sensing cantilever. Fig. 2.1b depicts these two situations. The system used in this work is in the latter configuration, which we consider to perform better since movements of the positioner are independent and have no effect on the cantilever unless there is direct contact. Nonetheless, our set-up is equipped with a second piezoelectric actuator, on the sensing part, which is capable of introducing vibration on the cantilever tip if needed (*e.g.* in tapping mode, see section 2.3). The positioners are equipped with sensing capacitors that measure the real position of the controller and correct it for hysteresis effects of the piezoelectric material when needed *via* a feedback mechanism.

It is important to notice that both positioning and sensing measurements are in volts. The calibration of the piezoelectric positioner is given by the manufacturer, but we recalibrate it in the laboratory using two independent methods: Imaging a ceramic sample with a step of known size and measuring the asymptotic length released after the unfolding of a canonical protein in the field – the 27th inmunoglobulin module from human cardiac titin (I27). The calibration of the cantilever sensor is done in each experiment using the thermal fluctuations method introduced in [53], as explained in appendix A.3.



(a) Sensing system



(b) Positioning system

Figure 2.1: AFM set-up. (a) From left to right, close-up image of the sensing cantilevers used in AFM, cartoon depicting a situation where the cantilever is not sensing a force, cartoon depicting a situation where the cantilever is sensing a negative force, such as the case where a protein is pulling from it. (b) Representation of the piezoelectric positioner (in red) in the situations where the movement is carried out on the sample (left), on the sensor (middle) or both (right). The system used in this work uses the last conformation, by which the positioning is controlled by moving the sample, but a secondary piezoelectric actuator is available if one needs to move the cantilever (*e.g.* in the tapping mode for imaging).

2.3. Imaging mode

Since AFM was invented as an imaging tool, it makes sense to start by reviewing its imaging capabilities. It was firstly used in only one configuration, but soon two other imaging modes were developed. Each of the methods presents its own advantages and handicaps, so that each of them can be used depending on the sample under study.

The first mode ever used was the contact mode [49]. It consists in approaching the cantilever tip and the surface until a specific contact force is reached. This force is known as set-point. Subsequently, the tip is dragged along the surface either at a constant height and registering the force, or changing the height so that the force remains constant. Either method yields excellent quality results for any surrounding medium, including vacuum, air or buffer, its only limitation being the size of the cantilever tip. However, if the sample is soft or not completely attached to the surface, dragging the sensing cantilever on it with a high force involves both indenting and shearing forces, which might damage or detach it. Since most biological samples are soft, less aggressive methods were needed.

One of the alternatives to this method is known as tapping mode, and goes also by non-contact mode or dynamic mode [54, 55]. In this protocol, a second piezoelectric actuator is needed (see Fig. 2.1b – right) in order to excite the cantilever and induce a controlled-frequency vibration. This method is based in the relation established between the amplitude of the induced oscillation and the distance from the cantilever tip to the sample, and can again be used in two different ways: maintaining a constant amplitude by changing the height or exploring at a constant height and measuring the amplitude. As a result, this method produces not only a topographic map of the surface obtained from the oscillation amplitude, but also provides a phase lag diagram from which viscosity, elasticity and other properties of the sample can be extracted. This method, being the least invasive for it involves no direct interaction with the samples, is very interesting to apply on biological samples. However, if the images are taken in liquid, the oscillation is considerably dampened and, therefore, the image quality is severely affected.

Eventually, a third method was invented based on SMFS experiments: the jumping mode [56, 57]. This protocol, also called intermittent-contact mode, consists in doing repeated cycles of approaching the sensor to the sample until it makes contact, then retracting and moving laterally when not in contact. It can also be carried out in two ways: measuring the height at which a specific contact

force is reached, or measuring the force at a specific height. This method implies direct contact with the sample in the normal direction, but does not involve shearing forces that may detach the samples from the substrate. Furthermore, indentation does not need to reach high forces as in contact mode, so deformation of soft samples will be small, if any. Finally, since vibration is not involved in this mode, it is ideal for liquid media.

Imaging AFM is a good methodology to study the morphology of a protein. Even without atomic resolution, the size and shape of a protein can be assessed using this methodology, as well as interactions if they arise. Furthermore, a method for *quasi*-simultaneous imaging and pulling was developed [58] where a molecule was first detected using an imaging tool and then the preferred spectroscopy protocol was applied to study its mechanical properties. Nonetheless, imaging typically involves very flat surfaces with non-specific binding interaction and low density of sample, while SMFS requires high amount of sample and strongly coordinated or covalent interaction of the protein with the substrate, therefore the combination of the two techniques is not trivial and, as for now, far from efficient.

2.4. Force spectroscopy mode

In the force spectroscopy mode of AFM, the molecule under study is attached to the tip of the cantilever and to the surface, and these two are pulled apart following a established protocol while measuring the force acting on the protein with the sensing cantilever and the displacement in the direction of movement using the capacitive sensors on the positioner. After the acquisition, and depending on the protocol used, the recorded data of force and distance are processed to obtain parameters such as the mechanical stability, the length released after unfolding, the elasticity of the molecule and kinetic parameters such as the unfolding and refolding rates and the distance to the transition state.

2.4.1. The Worm-Like Chain model

Proteins, from a reductionist point of view, are heteropolymers formed by twenty different kinds of beads (one corresponding to each residue). The nature of these beads typically leads the protein to form a secondary (local) and tertiary (global) structure, which might be force-resistant and thus observable in a force spectroscopy experiment. However, before and after the unfolding event, the protein

behaves like a polymer and it is therefore useful to use polymer physics to describe its behaviour.

The most typical model for polymer elasticity used on proteins is the Worm-Like Chain (WLC) model [59]. This model comprises a continuously flexible rod that presents correlation in the orientation of its segments in such a way that one tends to be aligned to the former one. In the case of a polymer subjected to a force, the model can be approximated by Eq. 2.4, that converges to the exact solution when $z/L_C \to 0^+$ and when $z/L_C \to 1^-$, and diverges up to 10 % (when $z/L_C = 0.5$).

$$\frac{F\,p}{k_B T} = \frac{1}{4(1-z/L_C)^2} - \frac{1}{4} + \frac{z}{L_C} \tag{2.4}$$

In Eq. 2.4, k_B is Boltzmann's constant, T is the absolute temperature, F is the force to which the polymer is subjected and z is the end-to-end distance. p and L_C are the two parameters that describe the polymer. The persistence length is the result of studying the correlation of the segments orientation, which decays exponentially along the chain distance with a factor $1/p$. It is thus related to the flexibility of the chain and, in the case of proteins, it has been measured to be around 0.42 ± 0.22 nm for globular proteins after unfolding. The contour length is the length of the polymer when stretched at infinite force. It is correlated to the number of residues that a protein has, and the slope has been measured to be 0.38 ± 0.18 nm per residue. The fact that these two parameters result in overlapping values leads to think that correlation is present between the residues in a protein after unfolding. However, results on proteins with no mechanical stability (or below the AFM resolution level of 10 pN) show that some peptides might present a persistence length of up to 2.5 nm [60], suggesting that there might be some cases that do show an interaction along the chain.

2.4.2. Single-molecule markers

The typical SMFS experiment in AFM involves contact between the sensing cantilever and the substrate where the sample is deposited on. This interaction is done at a high pressure: The cantilever radius is around 30 nm, and the contact force ranges between 1 and 2 nN, so the applied pressure is typically on the order of megapascals. Depending on many factors, including the material on which the sample is deposited, the sample concentration, its tendency to aggregate or the grade of purity of the sample and buffer, such a high pressure often results

in tip-substrate non-specific interactions detected as irregular behaviour at the beginning of a curve. Such behaviour is typically a random pattern and comes from non-specific sources unrelated to the specific sample under study. Partially or completely denatured molecules of the studied protein also contribute to the non-specific noise.

In order to discern the differences between data and non-specific noise, as well as make sure that the pulling is on only one molecule at a time and in the correct direction, single molecule markers were introduced. The effect of these markers is two-fold: They extend the experiment so that actual unfolding occurs far from the noisy region where the tip and the surface are close together (proximal region), and they provide an easy-to-detect pattern – typically periodic (or *pseudo*-periodic) – to mark the signal and distinguish it from the noise.

The first experiments carried out with AFM involved titin [51] and tenascin [61], two long modular proteins the modules of which have a similar characteristic shape. This served as single-molecule marker since the unfolding of one molecule would yield several similar unfolding events, most of them happening far from the proximal region. This technique, however, had two major limitations: One cannot use it unless the protein is modular and its modules are similar, and one cannot study the properties of each module – unless they differ in the length of the mechanically protected region – since the stochasticity in the unfolding process does not allow for peak-to-event correlation. Thus, alternative methods were developed to solve these problems, summarized in Fig. 2.2.

Homomeric polyproteins

The first strategy, based on the idea of these similar proteins, was developed in order to study one specific module. The method was called polyprotein strategy and was developed in two separate ways discerning in their making: One method is based on biochemistry [63] and another one involving genetic engineering [64].

The former consists in generating cystein mutations at specific points in the sequence of a protein module under study (which can be the whole protein), in order for it to bind to the next by a disulphide bond. The latter involves generating a recombinant DNA plasmid where one module is coded next to the other and express (make the bacteria generate the protein from the DNA) the whole recombinant protein as a single chain. An example of a polyprotein is depicted in the top panel of Fig. 2.2.

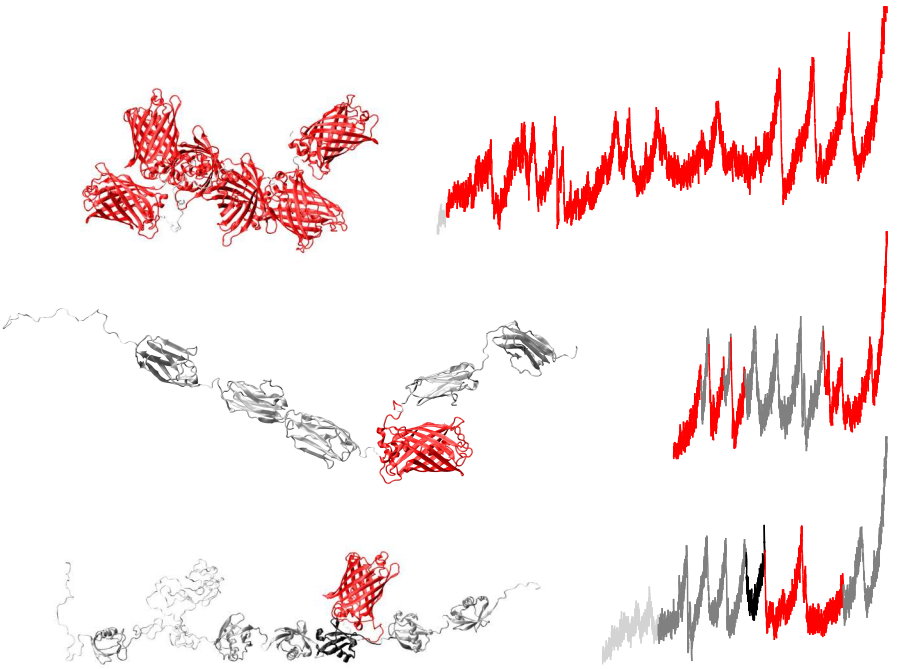


Figure 2.2: Single-molecule detection strategies. The top panel shows an example of a recombinant homomeric polyprotein of green fluorescent protein (GFP, PDB code 1GFL) in red with the unstructured linkers and tails in light gray. A force-extension plot is shown on the right where the colors match those in the protein. Using this technique with a protein that presents several unfolding paths such as GFP provides a bad single molecule marker, since periodicity is achieved only at the end while the beginning of the curve already contains partial unfolding events. The middle panel shows the heteromeric polyprotein strategy, using I27 (PDB code 1TIT) as marker in dark gray, GFP as the molecule to study in red and the disordered linkers and tails in light-gray. Using this method, the single-molecule markers can be readily identified, but the unfolding of GFP appears mixed with the signal of the markers and can be considered noise. Finally, the bottom panel depicts the Host-Guest strategy, using ubiquitin (PDB code 1UBQ) as single-molecule marker (dark-gray) as well as carrier (black), GFP as the molecule under study (red) and the disordered part of the molecule, including a fragment of titin's N2B region (light-gray). In this case, the signal is far from the proximal region and appears only after the carrier is unfolded. The figures of the proteins have been made using VMD [62].

The disulphide-bonding method has some advantages over the genetic-engineering one: one can choose from which points along the sequence to pull, the other method being restricted to pulling from the termini; and the production of such proteins is typically more efficient, since it depends on the size of the module to produce. However, it also presents several handicaps, such as having no control of the bonding efficiency and the final orientation of each of the modules (technical details which do not compromise since force is applied at both ends) but also this method needs oxidative red-ox conditions for the disulphide bonds to be formed and maintained, so studies involving reduction such as [65, 66] cannot be performed using this technique. Yet another advantage, if only technical, of genetic-engineering methodology is that the fusion proteins are expressed with a 6-histidine tag at the N-terminus used for affinity purification and two cysteine residues at the C-terminus to allow a covalent bond to a gold substrate – the most typical functionalization strategy for SMFS. Therefore, once expressed and purified, they can immediately be carried to the AFM to be stretched.

Apart of the lower performance in expression, the genetic-engineering method presents another handicap: the amount of work needed to generate a polyprotein DNA sequence starting with single-module ones. However, DNA lasts for long time once formed and is easily replicable, so once you have done a polyprotein, generating a new identical one is trivially accomplished by expressing the same clone again. This is why the genetic-engineering technique has become a standard in the field of SMFS [58, 67].

The use of polyproteins has been criticized in the past [68, 69], but an extensive comparative study on the well-known modules of ubiquitin and I27 in a monomeric form was carried out and validated the application of the technique for folded proteins [70].

Homomeric polyproteins represented a huge advance in the field. They are perfect single molecule markers, since the number of peaks exceeding the number of expected modules make you readily discard the trace for having more than one protein. They are also good for keeping the signal away from the noisy proximal region, since it can sacrifice the unfolding of one or two modules that will act as spacers at the beginning of the unfolding curve. But this technique not only solves the two problems we had with detection, it also adds two more very interesting features: To start with, each unfolding curve yields more than one unfolding event, meaning the data acquisition speed is multiplied by a factor depending on the number of modules the protein has. Secondly, it is capable

of enhancing effects that would otherwise be missed due to its low signal. An example of this was the discovery of the AB hydrogen-bonded region of the β-sandwich in I27, which results in a deviation from the WLC behaviour in a force-extension curve known as a hump [64].

Heteromeric polyproteins

Even if the use of homomeric polyproteins has been a great advance, the study of proteins with several unfolding pathways, more than one unfolding state or with a mechanical stability below the detection limit of the AFM ($\approx$ 20 pN) requires a specialization of this technique. Indeed, in a case with several unfolding pathways, unfolding events would be seen where the released length does not correspond to a complete module, breaking the periodicity of the registers as well as making it hard to associate each event with a module. As for low-stability modules such as PEVK and N2B regions of titin, they require some other single molecule marker, since the only feature that would appear in the registers is a long length.

To this end, the concept of heteromeric polyprotein was invented [60, 71]. In this case, a classic polyprotein of a known module (typically I27 or ubiquitin) is generated, and then one (or more) of the modules is changed to the new protein under study. In this case, the known modules act as single-molecule markers, while the unknown module is studied. An example molecule and force-extension trace is shown in the central panel of Fig. 2.2.

This technique preserves several of the advantages from its homomeric equivalent: on one hand the molecules retain the histidine purification tag and the cysteine residues for gold attachment, and single-molecule identification is assessed by the periodicity in the known part of the signal. It presents an important added advantage: the study of the single molecule markers yields an internal calibration control both for the force and for the length measurements.

Based on this technique, a ready-to-go expression vector[1] was manufactured, which includes eight tandem repeats of the I27 modules separated by different restriction sites [72]. This strategy makes it easier to generate a heteromeric polyprotein: Using specific restriction enzymes, specific I27 modules in the plasmid can be removed and substituted by the sequence corresponding to the protein

[1] A vector is typically a piece of circular DNA containing a bacterial cloning promoter and the sequence to be expressed.

under study. This protein can then be expressed and studied in the AFM without the need of building the DNA expression plasmid for the whole polyprotein from scratch.

This technique works fine with molecules with mechanical resistance to unfolding, especially when it is higher than that of the single-molecule marker, but in the case of proteins that tend to unfold before the marker it can be troublesome: the unfolding would then occur in the proximal region, and be mingled with the non-specific noise. Even if one selects the cleanest recordings, which further decreases the data acquisition efficiency, if the protein has several unfolding pathways its detection will still be tricky.

Host-Guest strategy

The hierarchical unfolding of the modules in a typical SMFS experiment, according to its mechanical stability, makes it difficult to study proteins that break at low forces because their unfolding events appear mixed with the non-specific noise in the proximal region. To overcome this problem, heteromeric polyproteins are not enough, the unfolding order needs to be inverted if one wants to study the unfolding of the low-stability module unequivocally. This can be achieved by using a mechanical protection strategy whereby the protein under study is cloned as a Guest (G) inside a more resistant protein, known as Host (H) [73]. In such a case, the H withstands the whole weight of the force and the G does not feel it until its protector gets unfolded. With this idea in mind, a new vector was designed [74]. This new vector includes five ubiquitin modules as single-molecule markers – which help to maintain the fold of proteins with chaperoning properties [75]. Moreover, it contains a fragment of the disordered domain N2B from titin, known to unfold with no detectable force [67], between the first and second ubiquitin markers (counting from the N-terminus). This disordered module was designed as a spacer to separate the unfolding of the actual modules from the noisy proximal region. Moreover, it is equipped with yet another ubiquitin module (for a total of six), known as the H and placed in position 4 from the N-terminus, which can contain, inside an insertion-tolerating loop, a G that will be protected from the force. An example of a recombinant protein following this strategy can be seen in the bottom panel of Fig. 2.2.

This vector was designed to study low-stability modules, but it has a wider field of application: Its use was crucial in the SMFS study of proteins involved in neurodegenerative diseases, which present a rich mechanical polymorphism

(see sec. 6).

2.4.3. Length-Control mode

AFM-based SMFS can be performed in two main operating modes. Length-Control (LC) mode is the one that was used in the first ever SMFS experiment [51]. It is often incorrectly called the Length-Clamp mode or the Force-Extension mode. The former comes as opposed to Force-Clamp (see sec. 2.4.4), which is not a correct name because the length is not clamped. The latter is because the information obtained is often displayed as a Force-Extension plot, which is also not correct since the analysis of an experiment should have nothing to do with its execution.

This is the simplest operation mode of AFM. The position of the piezoelectric positioner is continuously controlled, and any waveform is applied to it in order to achieve an approach-retraction cycle. The most typical function involves approaching and retracting at a constant speed. During the approach-retraction cycle, both position of the surface (z) and force (F) with respect to time are recorded. Before the collection of the data, an analog filter should be applied according to the Nyquist criterion [76]. Data treatment involves unit conversion from voltage to distance or force using the specific calibrations, the computation of the surface-cantilever distance (d) as the difference between the current position and the position where it makes contact (z_0) and the calculation of the end-to-end distance of the protein (ℓ), which is related to d but needs a correction for the cantilever bending. This correction is accomplished by assuming the cantilever is a linear spring and computing the corresponding expected position with time due to the measured force acting on it, as in equation 2.6, where k is the spring constant of the cantilever.

$$d(t) = z(t) - z_0 \tag{2.5}$$

$$\ell(t) = d(t) - \frac{F(t)}{k} \tag{2.6}$$

Eventually, the results are typically depicted as a $F(\ell)$ plot, where each unfolding event is marked by a force peak and, in the case of a modular protein like titin or a homomeric polyprotein, results in a sawtooth pattern.

Typical speeds of this experiments range from 10 to 2500 nm/s. Faster speeds introduce noise in the curves and yield the forces unavailable, while for lower

speeds the protein does not remain attached between the surface and the tip for long enough [77].

On plotting F *vs.* ℓ, one can fit a polymer elasticity model such as the WLC (see sec. 2.4.1) to each of the force peaks, which yields the elasticity of the chain as a persistence length and the contour length of the protein before unfolding. After fitting several such curves to the peaks of a sawtooth pattern, the difference between the contour length assigned to two consecutive peaks can be measured and correlates with the number of residues in the protein that are hidden to force. The distribution of such lengths for a globular protein is narrow, typically around 1 %, making this measurement a good identifier for each module.

Furthermore, the height of the peak marking each unfolding event corresponds to the force at which a molecule unfolded. This force is typically characteristic for each protein. In particular, the force corresponding to the highest peak is known as mechanical stability. The distribution of forces, however, is broader than that of contour lengths – between 10 and 20 %. This bigger dispersion is attributed to several factors, including the vibration of the cantilever, but also the intrinsic stochastic nature of the unfolding process, which mainly depends on the solvent accessing the force-resisting bonds.

Typical AFM experiments in the LC mode ultimately compute the intrinsic unfolding rate of a protein by performing repeated measurements of the unfolding force F as a function of the pulling velocity v and extrapolating to $v = 0$ using the logarithmic dependence from [78]. Nonetheless, even if LC-mode experiments are common, measurement of the unfolding rate using this protocol is costly and, therefore, rare.

Importantly, the logarithmic dependence has been recently proved to perform poorly for high speeds, and that the relation $\langle F \rangle \approx \ln^{2/3} v$, theoretically derived in Ref. [79], should be applied instead. This has been recently confirmed using a high-speed AFM to be able to pull at speeds up to three orders of magnitude higher than the top limit of regular AFM [80].

Folding experiments have been carried out in the LC mode with the goal of monitoring the active force of the unfolded polymer to refold. In this case, each attempt consists of touching the surface, pulling until the complete extension of the protein, and relaxing the tension again before the protein is detached from the substrate or the cantilever. Repeating this approach-retraction cycle, a protein can be unfolded and refolded several times. However, the folding force has not been measured in AFM using this mode, due to the presence of drift intro-

ducing force artifacts difficult to discriminate from real events [81]. Even so, the folding kinetics of some proteins have been studied using this protocol, including membrane proteins and solenoids [82, 83].

2.4.4. Force-Control mode

Complementary to the LC mode, the Force-Control (FC) mode performs the experiments exerting a control on the force. They were initially performed on I27 [84] and christened Force-Clamp mode, which is only a particular case of what can be achieved with this method. The main advantage of this protocol over LC is that time is one of the variables at play, so dynamic information can be directly extracted from the experiment. Another advantage of this technique is that it controls an intensive variable – *i.e.* a variable that does not depend on the size of the system – and thus it can be applied to several different proteins in a comparable manner, with results that are not experiment-dependent.

This method is more complex than LC. The AFM is a set-up designed to control the tip-surface distance and measure the force, so the control of the force needs to be applied through an extension or contraction of the distance. This is accomplished using a feedback mechanism which, if active, increases the surface-tip distance until the desired value of a force is achieved. The fact that a feedback mechanism is needed to use this mode can be troublesome for technical reasons. To begin with, the feedback mechanism needs time to correct the force value in the case of a sudden change – such as the unfolding of a module. This lag time is typically on the order of milliseconds, which depending on the pulling force can be more than enough time to unfold a second module, the unfolding of which would not be recorded independently of the first but as a single, longer unfolding event. Moreover, the feedback mechanism involves gains that control the behavior of the positioner. These gains being too low would make the system unable to retract enough to reach the desired force; while it being too high would amplify the noise in the force signal, due to real thermal motion of the sensor or artificial electric noise from the detection system, and induce a high frequency vibration movement of the positioner that would damage its properties. Nonetheless, adjusting the gain correctly one can easily perform experiments in the range of 10 to 200 pN for typical proteins.

Similar to the LC mode, any waveform can be supplied to the feedback mechanism for the force to follow, but the most typical ones are keeping the force constant, changing it in steps or changing it linearly with time. The first two

are commonly called Force Clamp experiments, while the last one is known as Force Ramp. In these experiments, force (F) and position of the sample (z) are recorded as a function of time as in LC mode, and they are treated similarly so that the end-to-end distance of the protein, ℓ, is obtained as in Eq. 2.6, while the plotting of the results differs.

In the constant and stepped force versions of FC, ℓ and F are plotted independently *vs.* time, and unfolding events are shown as steps in $\ell(t)$ and spikes in $F(t)$. If performed on a homomeric polyprotein, the $\ell(t)$ graph is like a staircase with several steps with the same rise. The spikes in $F(t)$ correspond to the feedback mechanism adjusting the force, and their width is the delay of the feedback mechanism, which ranges from 1 to 20 ms depending mainly on the gain used for the feedback, but also on the design of the device. In this case, the rise of the steps is also directly related to the number of residues hidden to force in the unfolded module, but contrary to the contour length, the rise of the steps is force dependent and can be obtained from the WLC model, Eq. 2.4.

Assuming that unfolding is a two-state process that depends on the action of a denaturant, the probability of a protein unfolding at time t follows an exponential law with an unfolding rate that depends on the pulling force. $\ell(t)$ plots show the complete extension of a protein with time, which increases with each unfolding event. Averaging several curves and fitting an exponential saturation to the total unfolded length yields the unfolding rate, and repeating the experiment for several forces results in having the unfolding rate as a function of the force, which can be extrapolated to $F = 0$ to know the intrinsic unfolding rate of the protein under study.

In the case of ramped force experiments, force and time become linked to one another and a $\ell(F)$ function can be trivially studied. The same two-state approximation results in ℓ being a double exponential on the force, from which the unfolding rate at zero force can be directly measured without having to repeat the experiment at different forces. Even if this experiment performs faster than constant force, the identification of correct events with the single-molecule markers is more difficult, since it presents steps of different sizes even for homomeric polyproteins depending on the force at which the unfolding occurred.

The FC protocol has been widely used not only to study protein unfolding but also to study protein refolding [85], the effect of the solvent in protein unfolding and refolding [86] and the effect of force on disulphide-bond reduction [87], among others. Recently, experiments have been carried out in order to find the

origin of the folding barrier and the shape of the energy landscape, leading to controversial results [88, 89, 90, 81, 91, 92].

2.4.5. Other force spectroscopy techniques

As aforementioned, AFM has a high spatial resolution (below 1 nm) but low force resolution (*ca.* 10 pN). Nonetheless, other single molecule manipulation techniques can be used to perform SMFS in complementary ranges. Among them, optical and magnetic tweezers are the most well known [93].

Optical tweezers have a force range of 0.1 up to 100 pN, while the typical spatial resolution is *ca.* 10 nm. The principle behind its working involves the gradient of light pressure induced on the focus of a laser beam, which is applied to a microscopic transparent bead located at its focus. It is this gradient that exerts a pressure on the bead and allows to move it to the desired positions. A typical optical tweezers experiment involves coating one bead with the sample under study, tagged at the free end with a molecule from an interacting pair (typically biotin). This bead is fixed at place using a pipette. Next, another bead coated with the other molecule from the pair (streptavidin) is trapped using the laser beam and approached to the first bead in order for a sample to attach. Finally, pulling experiments similar to those described for AFM are carried out. Optical tweezers have been used to study molecular motors and the properties of RNA and DNA.

Magnetic tweezers are based on a similar principle, but in this case the bead is magnetized and the laser is substituted by a magnetic field. The field ensures that the bead tends to move toward the highest intensity, and thus its movement can be directly controlled by moving the magnets. The force resolution is, in this case, much higher than that of AFM and optical tweezers, down to 10^{-2} pN, but then again, the maximum forces it reaches are on the order of 10 pN (depending on the size of the bead). Another advantage is that the magnets can be rotated and thus induce a torsional force, which cannot be accomplished with the former two techniques. Finally, the FC mechanism can be directly established without the need of a feedback mechanism, since the position of the magnet ensures a constant force being applied to the bead. The main handicap is that they have a much smaller spatial resolution, down to 20 nm. Among other uses, torsional properties of molecules have been investigated with this technique.

Specifics of our AFM set-up, which was originally designed for SMFS and imaging capabilities were added afterwards, are explained in ref. [94]. Details

on the experimentation protocol can be seen in appendix A.

2.5. Summary

To sum up, we have presented the experimental set-up of an AFM which will be used to acquire information on SMFS of neurotoxic proteins – particularly Q_n and Aβ. Its main modes including imaging, LC and FC have been explained and it has been put in context compared to other similar techniques.

3. *In silico* Experiments

The term *in silico* was first used in 1989 in reference to other Latin expressions used in biology, such as *in vivo* or *in vitro*. It is used to indicate that a certain laboratory experiment or natural phenomenon is simulated in a computer. These simulations can be performed in several ways, including Monte Carlo and Molecular Dynamics (MD). In this work, we are centered in the latter.

3.1. Molecular Dynamics simulations

Modelling a single object moving according to a force field is fairly simple and can often be solved analytically. A two-body problem is more complicated, but in some cases can be exactly solved. Nonetheless, three-or-more-body problems are, in general, unsolvable without the help of numerical simulations. Interestingly, biological systems such as proteins cannot be reduced to one or two bodies: They include several amino acid residues (around 100), each composed of at least 6 atoms, and often the water molecules around them in a box that is big enough for the protein not to feel the edges. That amounts to around 40 000 atoms in a system that interacts with one another in several non-trivial ways.

MD was proposed within the frame of theoretical physics in the late 1950s as

a method to solve many-body problems [95]. According to this model, each particle is treated as a body with some size, and interactions between them depends on the potential energy of the system [96]. This energy is, in general, treated as the sum of several terms, each depending on a different number of particles, as shown in equation 3.1.

$$V = \sum_i v_1(\vec{r}_i) + \sum_i \sum_{j>i} v_2(\vec{r}_i, \vec{r}_j) + \sum_i \sum_{j>i} \sum_{k>j} v_3(\vec{r}_i, \vec{r}_j, \vec{r}_k) + \cdots \tag{3.1}$$

In this equation, V stands for the whole potential, while each term in the sum is v_α. Latin indices i, j and k run over the number of particles. The notation $\sum_{j>i}$ indicates that the interaction is only counted once for each pair of atoms, since the interaction for i–j is the same as the one for j–i.

Thus, the first term of the potential ($v_1(\vec{r}_i)$) only depends on the position of each particle and thus accounts for the presence of fields not generated inside the system. The rest of the terms detail particle–particle interactions. In particular, the second term depends on the position of pairs of atoms, which can be reduced to the relative position between each pair: $v_2(\vec{r}_i, \vec{r}_j) = v_2(\vec{r}_i - \vec{r}_j) = v_2(\vec{r}_{ij})$. Some of these interactions are, for example, Coulombic potentials in the case of ions.

Initial simulations carried out in solids and liquids assumed that more-than-three-particle interactions were small compared to the other terms and were discarded [96]. Furthermore, it is known that computing n-body interactions scales exponentially with n, so in order to significantly reduce the computation time three-body interactions were approximated by an effective two-body potential. Due to this approximation, even if the original potential is only position-dependent, the effective potential used in depends on other parameters such as temperature or density.

The interaction between particles can be modelled in many ways. The initial model is the ideal gas approximation, where each of the particles of the system is unaware of its neighbours. This is a simplification only useful for low densities and high temperatures, and the statistics of these simulations compare perfectly to the statistical mechanics of an ideal gas. Other models include the hard-spheres potential, where the collision of two spheres is prevented; the square-well potential, where there is a forbidden followed by an encouraged range of distances; and the soft-spheres potential, where the collision is disfavored as a (typically exponential) function of the distance between the two. Nonetheless, the most typically used potential function for the interaction between two par-

ticles is known as the Lennard-Jones 6–12 potential (Eq. 3.2), which is often empirically determined.

$$v(r) = 4\varepsilon \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right) \tag{3.2}$$

Equation 3.2 behaves as a long-range attraction combined with a steep repulsion, resulting in an energetically favourable position. Two parameters can be adjusted to the behaviour of each system: σ determines the position of the energy minimum, *i.e.* the equilibrium distance for the system; and ε is used to represent the strength of such interaction. This potential was successfully applied not only to gas systems but also in liquid simulations [96].

The study of molecules instead of free atoms is a more complex matter. Ideally one would like to use realistic quantum-mechanical force fields in order to model the bonds between atoms, so that the electronic density for each is correctly accounted for. However, integration of quantum-mechanical systems with so many degrees of freedom carries an intrinsically heavy computational load. To avoid this problem, the bonds were modelled as extra terms in the potential function [97]. This terms, however, are often not simply distance-dependent but involve also angular terms, so in the case of molecules, typical potentials include closest-neighbours interactions, meaning that covalent bonds are expressed in the potential as hard conditions on distances between each pair, bond angles between each triplet and torsion angles between each four covalently bound atoms. Therefore, the canonical potential function used in MD includes a bonded term and a non-bonded term, the former taking care of the distance between first neighbours, angles between second neighbours and dihedrals between third neighbours; and the latter considering only interactions of a non-covalent nature, such as van der Waals or electrostatic.

Classical mechanics provide us with a simple way of computing the equations of motion of a system once its potential and kinetic energies are known [98]. The typical set of coordinates and velocities chosen for MD simulations are the positions and velocities of each of the centers of the atoms, which reduces the equations of motion to Newton's second law. Thus, the movement of the particles described as a function of time are the result of the integration of N coupled differential equations, where N is the number of atoms in the system.

Free dynamics studies the behaviour of a system that involves no specific external force. After some equilibration, this simulation yields a model of a

protein in equilibrium. It can be used to study chemical denaturation through temperature or pressure, as well as conformational changes that may occur in fast timescales.

Steered dynamics refers to the use of a driving force on an atom (or a set of atoms) in order to achieve some result. It is typical to replicate SMFS experiments *in silico* with this methodology to see what contacts are relevant to each unfolding event. In particular, constant velocity pulling simulations similar to the typical LC experiment apply two restrictions to two atoms: one (typically at the N-terminus C_α) is an elastic restriction that restrains the movement of the atom, while the other (typically at the C-terminus C_α) carries the other atom away at a constant velocity. Also, equivalent to FC experiments, a constant force can be applied between the two atoms to study the time it takes for them to unfold.

MD has many good points, some of them being the access to all the information of the system, the low cost compared to experiments and the versatility of the system – in that all parameters can be controlled. Furthermore, in the field of SMFS, they have the additional advantage of being directly comparable to experiments because the studies are typically performed on a single molecule. Nonetheless, one must always take into account that MD are models of the reality and reach as far as the model goes, and that the timescales available in simulation are typically in the order of μs, even if some recent computing has been able to reach 1 ms [99].

The first application of MD to SMFS was published one year after the first pulling experiments of titin [100]. This first simulations already revealed two interesting features of MD: they can explain what is happening in experiments and they can predict new features not yet discovered. In particular, they explained each of the unfolding peaks in the force-extension curve as the breaking of the hydrogen bonds between two parallel beta strands present at the ends of each of the immunoglobulin domains of titin, which was later called mechanical clamp, and predicted a smaller peak due to the breaking of other bonds present between another pair of strands in I27. The first polyprotein generated was able to amplify the latter and observe it as a hump in the WLC-behaviour (see sec. 2.4.2) [64]. Furthermore, recent simulations have proved wrong an experimental set-up, leaping beyond the predictions. Indeed, MD experiments proved that measurements of the unfolding force that a proteasome needs to unfold a protein were measured experimentally with an incorrect configuration whereby the forces can be highly over- or underestimated [101].

3.2. Bias Exchange Molecular Dynamics

As stated in the first chapter, IDPs are rapidly fluctuating proteins that present scarce conformations with an ordered state. Furthermore, these uncommon species cannot be captured by high-resolution experimental techniques such as X-ray crystallography or NMR due to their inherent volatility. Therefore, MD seems a more than appropriate method to obtain atomic-resolution structure of these systems.

However, the exploration of a wide-enough landscape in the case of IDPs would require simulation times of the order of seconds, way above the current limit of microseconds. To that end, we used a method known as Bias Exchange Molecular Dynamics (BEMD) [102], designed as a combination of two others: Replica Exchange MD and metadynamics.

3.2.1. Replica Exchange Molecular Dynamics

Replica Exchange MD is a method that was conceived to speed up the exploration of an energy landscape [103]. The idea behind this method is to copy the system several times, each called a *replica*, and perform an independent simulation of each replica at a different temperature. Then, after some simulation time, replicas i and j are allowed to switch from one temperature to the other with exchange probability given by Eq. 3.3, where T_i and T_j are the temperatures at which each replica is being simulated, k_B is Boltzmann's constant and E_i and E_j are the energies at which each of the replicas is.

$$P(i \leftrightarrow j) = \min\left(\exp\left(\left(\frac{1}{k_\mathrm{B}T_i} - \frac{1}{k_\mathrm{B}T_j}\right)(E_i - E_j)\right), 1\right) \tag{3.3}$$

This method allows each replica to explore in detail every energy minimum when at low temperature, and to escape local equilibrium states with high barriers when at high temperatures, so the exploration of the energy landscape is performed thoroughly and at a great speed.

An extension to this method, known as Hamiltonian Replica Exchange, allows the system to be simulated not at different temperatures, but under different energy functions. This method in combination with the self-learning Hamiltonians explained in the next subsection will yield the BEMD approach.

3.2.2. Metadynamics

The metadynamics approach was first described in 2002 [104]. It is efficient in thoroughly exploring the energy landscape of a system along a specific reaction coordinate in a small amount of time. The chosen reaction coordinate is typically represented by one or more observables, each called collective variable, which need to be differentiable quantities.

Once a collective variable is chosen, the potential energy of the system is continuously modified by adding a memory term discouraging those conformations where the collective variable value has already been visited. These are typically Gaussian terms with a width and amplitude chosen in a compromise between efficiency (larger) and accuracy (smaller), such as the one presented in Eq. 3.4, where the collective variable is represented as S.

$$V(\vec{r}_i, t + \mathrm{d}t) = V(\vec{r}, t) + A \exp\left(- \frac{\left(S(\vec{r}_i, t + \mathrm{d}t) - S(\vec{r}_i, t = 0)\right)^2}{2\sigma^2} \right) \tag{3.4}$$

The efficiency of this metadynamics approach is measured as the number of Gaussians needed to fill the free energy landscape, which is proportional to the inverse of their width. Nonetheless, the width needs to be much smaller than the length scale of variation of the energy landscape in order to explore the landscape accurately. Furthermore, the addition of n more collective variables will increase the dimensionality of the problem n-fold, leading to an efficiency that scales exponentially with n. Thus, while exploring the energy landscape along one collective variable might be fast enough, having several of them is extremely time-consuming.

In this sense, the use of a Replica-Exchange-like methodology enhances the efficiency of the metadynamics approach.

3.2.3. Bias Exchange Molecular Dynamics

BEMD is a technique that wisely combines Replica Exchange with Metadynamics. Two methodologies can be used, as follows: The first approach involves simulating several replicas at different temperatures, each biased along a different collective variable, but updating the energy landscape of each replica collectively with the Gaussians generated not only in that same replica but also in the rest of them [105]. This method was shown to perform more than three times

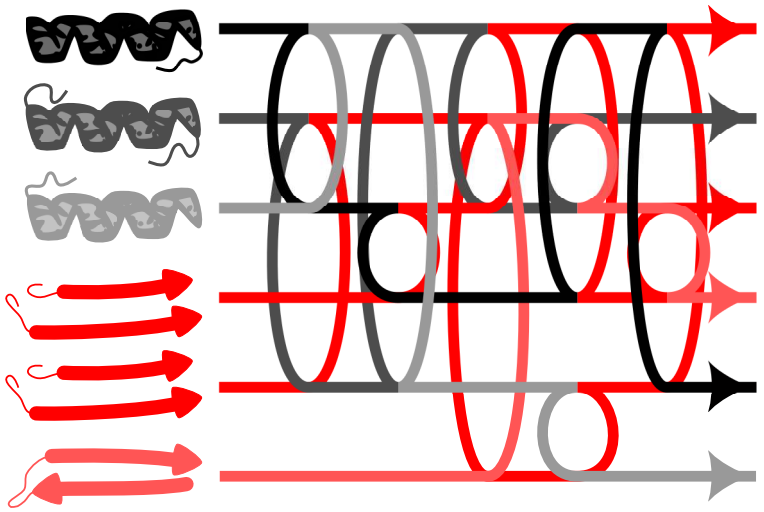


Figure 3.1: Bias Exchange Molecular Dynamics. The method consists in simulating several replicas of the system in parallel biasing each of them with a different potential function, and allowing exchange every some steps. In this work, six replicas were used: three were biased with α-helical structure at the first, second and last thirds of the protein – shades of black –, two with parallel β-strand bias – darker red – and one with antiparallel β-strand bias – lighter red.

faster than Replica Exchange MD in the exploration of a free energy landscape of a short protein, and the changing temperatures allows for the biases to flatten the energy landscape more effectively.

The alternative, which is the one we use in the work presented here, is depicted in Fig. 3.1. It is based on simulating several replicas of the original system parallelly and at the same temperature, each biased along a different reaction coordinate [102]. Each bias is then allowed to be exchanged from one replica to the other, thus affecting not only the system it was initially on, but also the rest of them. After enough exchange steps, all the replicas have suffered the effect of every bias, which has lead them to explore the whole energy landscape. Similar to the previous methodology, the exchanges allow a fast exploration of the whole landscape while the biasing allows it to be thorough.

In some cases, it is useful to add a neutral replica, *i.e.* a replica which presents no bias. This replica is useful because the sampling is, in this case, the real free

energy landscape of the system instead of a projection of this landscape on one of the reaction coordinates related to the collective variables. In the case of IDPs, however, the free energy landscape is extremely flat and it is preferable to skip this replica in favor of adding another bias for a quicker sampling.

3.2.4. Collective variables

One of the key points of BEMD simulations is to choose the correct collective variables. The ideal ones to chose in a folding simulation are each of the Ramachandran angles in the protein, thus being able to test the exact fold of each of the conformations. Nonetheless, even for a small (16-residue) protein, this leads to a very large number of collective variables (30), and therefore an impractical number for replicas. Thus, other collective variables need to be chosen.

In our case the goal is the acquisition of secondary structure, which leads to choosing as collective variable the presence or absence of a specific secondary structure group. In particular, the biases we chose for this work are three: Presence of α-helix, presence of parallel β-strands and presence of antiparallel β-strand. Nonetheless, "presence of" something is not a good definition for a collective variable, since the collective variable needs to be differentiable in order to be added correctly to a potential function and therefore result into a driving force. To that end, we use the collective variables suggested in Ref. [106], which are explained next.

Let us take, as an example, the case of an antiparallel β-strand collective variable. In order to study the current conformation of the atoms in the molecule regarding this collective variable, we group the residues in the system in the form $\{i, i+1, i+2, i+h+2, i+h+1, i+h\}$. For each possible group of residues of this form that can be generated in the protein, the backbone N, O and C as well as the C_α and C_β are selected, and the Root of the Mean Square Deviation (RMSD) of their positions is computed against the ideal 3+3 β structure. In order to obtain the ideal 3+3 β structure, the set of twenty protein representatives classified as "mainly beta" in the CATH database [107] are studied and the central structure is chosen to represent an ideal structure.

Once the RMSD is obtained, it is normalized in the function $n(x)$, which runs from 0 (for large x) to 1 (for small x) in a smooth way, so that differentiation is possible. This function is defined in Eq. 3.5, where x is measured in Ångströms.

$$n(x) = \frac{1-x^8}{1-x^{12}} \tag{3.5}$$

Eventually, after $n(x)$ has been computed for all possible subgroups of the specified form in the protein, the sum of all of them yields the total collective variable for the system.

This methodology can be readily generalized to parallel β-strands by taking the sets of the form $\{i, i+1, i+2, i+h, i+h+1, i+h+2\}$. Equivalently, it can be extended to α-helices with $\{i, i+1, i+2, i+3, i+4, i+5\}$. The determination of the differentiable collective variable is then achieved by applying the same method.

3.3. Structure-based Molecular Dynamics

All-atom MD used to study a single protein of average size involves the simulation of tens of thousands of atoms. Even if this is becoming more and more available with the technological advances, the reality is that simulations longer than tens of nanoseconds are rare exceptions, even if these can reach the millisecond scale in some specific cases. In that spirit, and since many biological processes that occur at the molecular level take longer than 10 ns, MD simulations need to be simplified in order to achieve relevant timescales.

Simplifications can be done in many places, although maybe one of the most typical ones involves substituting the solvent by a term in the equations of motion representing Brownian motion in terms of random forces that temper the system to a specific temperature. Nonetheless, even this reduction might in some cases be not enough simplification, especially if a large number of tests need to be carried out. To that end, Gō and collaborators proposed in 1981 a structure-based model in the context of protein folding and unfolding [108].

3.3.1. The model

Structure-based modelling consists in substituting groups of atoms by a single entity, then changing the potential to an effective potential that controls the behaviour of the group. The simplest approach is to represent each amino acid residue in the protein by a bead of a specific size at the location of the C_α atom of the residue, then tether these beads along the peptidic chain using harmonic

potentials with an equilibrium distance equal to that of the experimentally obtained structure. The beads are typically given a radius of 0.4 nm, thus yielding a soft excluded volume effect and avoiding the self-crossing of the backbone – a non-realistic movement.

These particles, being bigger than typical atoms, are subjected to a diffusive type of dynamics, which induce an intrinsic difference between this kind of simulation and the traditional all-atom one. While in the latter the characteristic times are determined by the ballistic movements and are therefore of the order of 1 ps [96], in a coarse-grained model the motion is mainly diffusive, and thus the time scale, τ, is of the order of 1 ns [109].

The key part of structure-based modelling is distinguishing the native contacts – those that are formed in the native state – from the non-native ones. A list of the native contacts, known as the contact map, is determined from the experimental structure and the simulation considers those contacts to be attractive, as opposed to the non-native ones. The method for determination of the contact map, as well as the specific form of the attractive potential and the stiffness of the peptidic bonds are model-specific, but a survey of 62 different models found one that outperformed the others when compared to experimental studies [110], which is explained next.

Backbone stiffness

The backbone stiffness is computed using a chirality-based approach, as opposed to an angular-based one, due to the higher computational complexity of the latter. The chirality potential for each atom quartet is given by Eq. 3.6, κ is a dimensionless parameter that controls the strength of the potential, ε is the energy paremter, ξ_i stands for the chirality of residue i, ξ_i^{N} for its chirality in the native state, $\vec{w}_i = \vec{r}_{i+1} - \vec{r}_i$ and $d_0 = |\vec{w}_i|$. As explained in Ref. [111], the value of κ needs to be selected. In this work, κ is taken to be 1.

$$V_\xi = \sum \frac{\kappa}{2} \varepsilon \left(\xi_i - \xi_i^{\mathrm{N}}\right)^2 \; ; \quad \xi_i = \frac{(\vec{w}_{i-1} \times \vec{w}_i) \cdot \vec{w}_{i+1}}{d_0^3} \tag{3.6}$$

Native contact energy

The potential energy of the native contacts is computed using a 6–12 Lennard-Jones potential as presented in Eq. 3.2, repeated here for convenience.

$$v(r) = 4\varepsilon \left(\left(\frac{\sigma_{ij}}{r} \right)^{12} - \left(\frac{\sigma_{ij}}{r} \right)^{6} \right) \tag{3.2}$$

In the case of structure-based dynamics, a σ_{ij} is needed for each native contact, and they are chosen so that the equilibrium distance is the original distance in the experimental structure. The energy parameter, ε, is chosen to be fixed for all contacts, even if other scenarios could be envisioned. One should notice that choosing ε fixed does not take away the inherent heterogeneity of the contacts – them being between different residues. This heterogeneity is preserved since it comes from two other sources: The σ_{ij} parameters chosen, which depend on the specific positions of the atoms (which in turn depend on the residue types); and the final contact map, which will vary due to the residue type as well.

Native contact determination

One other aspect remains to complete the model selection, and it is the contact map determination. In order to generate a contact map, we use the native structure and apply the following algorithm: Firstly, we select the heavy atoms in each residue. To each of them we assign a van der Waals radius, taken from ref. [112], which depends not only on the element but also on the bond it establishes with its neighbours. These radii are multiplied by $\alpha = 1.24$ in order to account for attraction. Next, all the spheres corresponding to each atom in a residue are grouped together and put in the context of the rest of the residues, equally enlarged. If there is an overlap between residues i and j, a contact is established between them. This method, known as overlap (OV) is depicted in Fig. 3.2.

The OV method, due to the expansion of the heavy atoms, usually presents contacts not only with the first neighbours, but also with the second ones (contacts of the type i, $i+2$). These contacts are already taken into account in the chirality potential, and therefore should not be considered in the contact map. Thus, the contact map is constructed using only contacts between residues that are three or more positions apart in the sequence.

3.3.2. Dynamics

Once the protein has been modelled, one can study several properties using this model, such as thermal and mechanical denaturation, or folding from a stretched

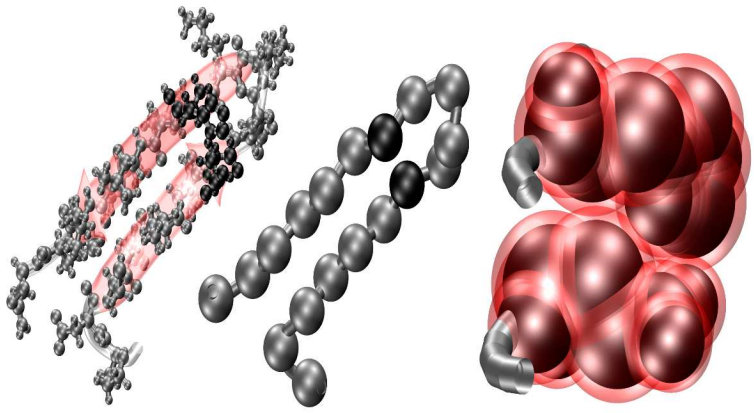


Figure 3.2: The OV contact-map-determination algorithm. The left panel represents the all-atom representation of a β-hairpin, on which the secondary structure is depicted and one of the contacting residues is highlighted in black. The central panel is the same β-hairpin, in this case simplified to a coarse grained scheme with a bead at the position of each C_α. The right panel shows a zoom of the previously highlighted residue, each atom with its corresponding van der Waals size, and an enlarged version of it in transparent red to account for attraction. Since the two enlarged residues overlap, a contact is established.

conformation. In order to study all these properties, the system needs to evolve in time. The evolution is achieved using an overdamped medium approximation in a Newtonian dynamics as shown in Eq. 3.7.

$$m\ddot{\vec{r}}_i = -\gamma\dot{\vec{r}}_i + \vec{F}_i + \vec{\Gamma}_i \tag{3.7}$$

In this equation, m stands for the mass of the particles, which has no effect in overdamped dynamics and is thus chosen to be 1 for all residues [113], γ is the damping coefficient, which has been determined to be optimal at $2m/\tau$ [109, 113]; $\vec{F}_i$ is the deterministic part of the force accounting for the potential function; and $\vec{\Gamma}_i$ is the stochastic part of the force, which takes care of the tempering of the system by introducing random forces normally distributed around zero with width $\sqrt{2\gamma k_{\mathrm{B}}T}$.

Once the dynamics have been implemented, the study of thermal unfolding proceeds by dynamically changing the temperature and monitoring the number of formed contacts. Contacts are determined to be broken once the contact distance is greater than 1.5σ, well over the inflection point of the Lennard-Jones potential, where the energy is around 30 % of that of the minimum.

Similarly, in the case of mechanical unfolding, the breaking of the contacts is monitored as a force is applied. Both protocols explained for the AFM (see Sec. 2) can be used here: one bead, typically the N-terminal one, is kept in place while another, usually the C-terminal one, is either moved away at a constant speed (LC) or subjected to a constant force (FC). Both the static and the dynamic constrains are both applied using an elastic spring, and the simplicity of these models allows for the pulling to be performed at $5 \cdot 10^5$ nm/s, three orders of magnitude over the typical experimental scale of 500 nm/s, whereas using all-atom simulations the stretching needs to be carried out seven orders of magnitude faster than the experiments. By comparison to stretching experiments, the value of the ε term of the potential has been calibrated to be (11 ± 3) pN nm.

For protein folding, the protein typically starts in an extended conformation and is allowed to evolve under a potential determined by the contact map. in this case, instead of determining when the contacts are broken, one needs to look at the time when the contacts are being formed.

An interesting representation to look at both in unfolding and folding simulations is the scenario diagram, a plot where the breaking or the formation time of each contact is plotted *vs.* an identifier such as the sequence distance between

the two interacting residues. This plot gives an idea whether the amount of contacts being broken is big or small, or if the contacts being formed are between nearby residues or long-range interactions, among other information.

3.4. Specifics on Molecular Dynamics simulations

After introducing the *in silico* techniques used in this work, I will explain the details we used in the simulations, such as parameters and programs.

All-atom simulations were carried out using GROMACS 4.6 [114] simulation suite with the AMBER99 [115] force field. Structures were obtained from the PDB when available, or were modelled on- or off-template using the MODELLER software [116]. The simulations in this work were carried out using implicit solvent with the generalized Born surface area model [117], so no water box was added to the system.

After the structure was obtained, a two-step minimization process was carried out: Firstly, the molecule was drawn to a minimum using steepest descent until the maximum force present between a pair of atoms was smaller than 0.25 J/(mol nm) (or for no longer than 10 000 steps). After that, a finer approach to the minimum was carried out using the conjugate gradient method also until the maximum force was smaller than 0.25 J/(mol nm), but this time the system was allowed to evolve for 40 000 steps. During the minimization steps, all C_α atoms are held in place by elastic restrictions with high spring constant ($10\ \mathrm{MJ}/(\mathrm{mol\ nm}^2)$). In the cases where the initial structures were generated instead of directly obtained from the PDB, 10 different models were created. From these, the knotted ones (if any) were discarded and the rest were minimized as formerly explained. Only the one with the smallest potential energy among them was chosen as the initial structure to continue simulating.

After minimizing, the temperature of the system was risen in an equilibration process. Thermostating in this phase is done using the velocity-rescale method, since it converges fast and still yields a correct canonical environment [118]. During this process, the C_α atoms were still restrained. This process is typical for most all-atom simulations in order to accommodate the solvent molecules, if present, but also to allow for hydrogen atoms – which often need to be added artificially – to move slightly until they are actually at the minimum free energy state. The temperature rose at a rate of 25 K/ps doing restraint dynamics with a time-step of 1 fs until the desired temperature, typically 300 K. After temper-

ing, the system was stabilized by logarithmically reducing the strength of the restraints on the C_α's by a factor 2 every 2 ps down to 625 kJ/(mol nm^2), which we consider to be sufficiently small to remove completely.

Finally the protein is ready to be simulated, and free dynamics is run. The free dynamics is performed in the canonical ensemble using the Nosé-Hoover thermostat, a second order thermostat that preserves the canonical ensemble more efficiently [119, 120]. The time step in this case is 2 fs and the hydrogen atoms are constrained by the SHAKE algorithm [121].

In the case of BEMD simulations, the PLUMED 1.2 package [122] was added to the GROMACS suite. The Q_n system was simulated with six replicas, where three had an α bias, each on a different third of the protein sequence; another had an antiparallel β bias and the other two had parallel β biases. This is so because the exchanges between an even number of replicas favour an even exploration of the landscape, while the use of an odd number of them can lead to errors in the exploration (Cossio, personal communication). The Aβ case was simulated using four replicas: Due to the smaller size, the α-helix bias was applied to the whole of the protein at the same time instead of the three thirds. The specifics of these biases are detailed in Sec. 3.2.4. The simulation temperature was 400 K, which yields a faster exploration of the energy landscape [123]. The biases are added to the potential in the form of Gaussian functions of height 20.92 kJ/mol and width 0.3 nm every 10 ps, and exchanges between biases are allowed every 25 ps. We save the coordinates of all atoms in the system every 5 ps, which generates a snapshot.

Finally, in the case of coarse-grained MD, the simulations were run using a code written by Prof. M. Cieplak and collaborators. The contact map is generated, as explained in Sec. 3.3.1, using the OV algorithm. The integration is performed using a fifth-order predictor-corrector algorithm [96] with a time step of 0.005 $\tau \approx 5$ ps). The temperature used was 0.3 ε/k_B, which is around room temperature. Furthermore, in constant velocity stretching, the protein ends were separated at a constant speed of $5 \cdot 10^{-3}$ Å/$\tau \approx 5 \cdot 10^5$ nm/s, while constant force simulations were performed for forces running from 5 to 30 ε/nm (55 to 330 pN).

3.5. Summary

This section has focused on discussing the methods for carrying out computer experiments, in particular the ones we are going to develop in this work. It explains

the details of the simulations as well as the algorithms involved. The methods explained here will be useful for the generation and study of the different conformers of Q_n and Aβ.

Overview of the thesis

This work focuses on different but related topics. It addresses technical issues both in the experimental and in the theoretical parts of SMFS, and addresses the issue of the study of IDPs related to disease from a single-molecule point of view. In particular, the objectives of the thesis are as follows.

1. Validation of the Host-Guest strategy as a categorical tool to study protein unfolding by AFM.
2. Inquire the relevance of the model for the contact map generation in the context of structure-based molecular dynamics simulations.
3. Study the mechanical properties of polyglutamine expansions and β-amyloid at the monomer level by SMFS.
4. Explore the conformational space of polyglutamine and β-amyloid *in silico*, also at the monomer level.
5. Characterize the resulting conformers in geometrical, structural and dynamical terms.
6. Study the toxicity mechanisms of the studied IDPs as unfolded through the proteasome.

Part III

Analysis

4. Exhaustive exploration of the Host-Guest strategy

The *H*-*G* strategy for single-molecule identification was proved to work experimentally at the time of its conception [73, 74], and has thereafter been used to study the behaviour of proteins with low mechanical stability ($F_{\max}$) or that present a polymorphic unfolding [124, 125], but a comprehensive study of the properties of the proteins that can be used with this technique was missing.

To address this issue, we performed MD simulations of several combinations of *H*s and *G*s, as well as single-molecule markers. In particular, it is interesting to know whether the effect of having one protein hidden to force inside another could result in differences in its unfolding pattern as compared to it being on its own. To generate these models, we embedded one protein inside the other with care that no contacts were formed between them. To that end, if the loop where the protein was inserted was not enough to provide separation, short alanine chains were added to act as spacers linking the *G* and the *H*.

MD played a key role in this study, since the fact that we would like to compare the unfolding of proteins with several ranges of $F_{\max}$ studied both in the *H*-*G* strategy and on their own makes AFM experiments labor-intensive. Fur-

Protein name	PDB code	F_{max} [ε/nm]	F_{max} [pN]
Barnase	1BNR	11	120
Ig-binding domain of protein G	1GB1	20	220
I27 domain of titin	1TIT	21	231
Ubiquitin	1UBQ	22	241
Cohesin module from *C. cellulolyticum*	1G1K	39	426
β domain of streptokinase	1C4P	51	562

Table 4.1: Isolated mechanical stability (F_{max}) of the proteins used in the study of the universality of the host-guest technique.

thermore, due to the system size, we decided on using structure-based MD (see sec. 3.3) to speed up the computations.

The study was carried out by comparing several cases: Those where the markers are the same as the host, with a guest with varying F_{max} (being lower, similar or higher than that of the host), and then cases where we change the F_{max} of the host compared to the markers. In order to explore a wide range of F_{max}, we used six proteins taken from the PDB, which had been previously analyzed at least with the same MD model, although most of them have also been studied experimentally. These proteins, ordered from lower to higher F_{max}, are Barnase (1BNR), Inmunoglobulin-binding domain of streptococcal protein G (1GB1), I27 domain of titin (1TIT), Ubiquitin (1UBQ), Cohesin module from *Clostridium cellulolyticum* (1G1K) and β domain of streptokinase (1C4P). Tab. 4.1 collect their isolated F_{max} as computed using the same structure-based model we used in this study [126].

4.1. On the mechanical stability of the Guest and the Host

In a first approach, we wanted to rule out the possibility that the mechanical properties of the G could be modified by it being in the H, or forming part of the complex. To do this, we studied five identical serially connected molecules, the central one acting as H and the other four as single-molecule Markers (Ms). The G was grafted in the H between positions i and $i+1$, leading to the configuration represented in 4.1. It should be noted that this is not the only possible scenario, but it is the simplest in order to maintain as much as possible the mechanical properties of the H.

$$2 \cdot M - H(i,G) - 2 \cdot M \tag{4.1}$$

In this part, we chose $M = H$ being 1UBQ, and G varying between 1BNR, 1UBQ and 1C4P so that the $F_{\max}$ of the G is smaller, equal and greater than that of the H. The results can be seen in Fig. 4.1. With this simulation we prove that the $F_{\max}$ of the G is not dependent on the H. Furthermore, this already hints at the conclusion we will get in the following section: The molecule of interest, the G, can be unequivocally identified by seeing the unfolding of the H before it; but it can be unclear since it may be mixed with the Ms. This could be solved by controlling the unfolding order, and making the H unfold after all the Ms have.

4.2. On the mechanical stability of the Markers and the Host

Next we want to study how the ratio between the $F_{\max}$ of the H and the Ms affects the unfolding trace. To this end, we use the same configuration described in Eq. 4.1, this time using different modules for H and M. In particular, the Ms are still 1UBQ, the G is 1TIT and the H changes between 1BNR (lower $F_{\max}$), 1G1K (higher $F_{\max}$ and two unfolding peaks) and 1C4P (higher $F_{\max}$ and one unfolding peak). The possibility of the H having a similar $F_{\max}$ as the Ms is already contemplated in the previous section. One can observe the results of this comparison in Fig. 4.2.

In this case we see that when the H has a small mechanical stability, its unfolding is the first of all the modules, and thus the unfolding of the G comes mixed with the Ms. Therefore, even if the signal is unequivocally detected and separated from proximal noise, it is still difficult to analyze, especially in the case of a polymorphic G.

In the case of the two-peaked H the situation is similar: The signal for the H comes after all the Ms, but since the height of the second peak of the H is similar to the $F_{\max}$ of the G, the signals can come mixed and the interpretation of the signal, although still unequivocal thanks to the contour length release after each peak, is not direct.

Finally, the case where the H has a high $F_{\max}$ and unfolds in a single step is the preferred case: The single-molecule markers provide a fingerprint to assess we are working with a single molecule instead of many, the unfolding of the H ensures the pulling force is applied on the G at the N- and C-termini, and its

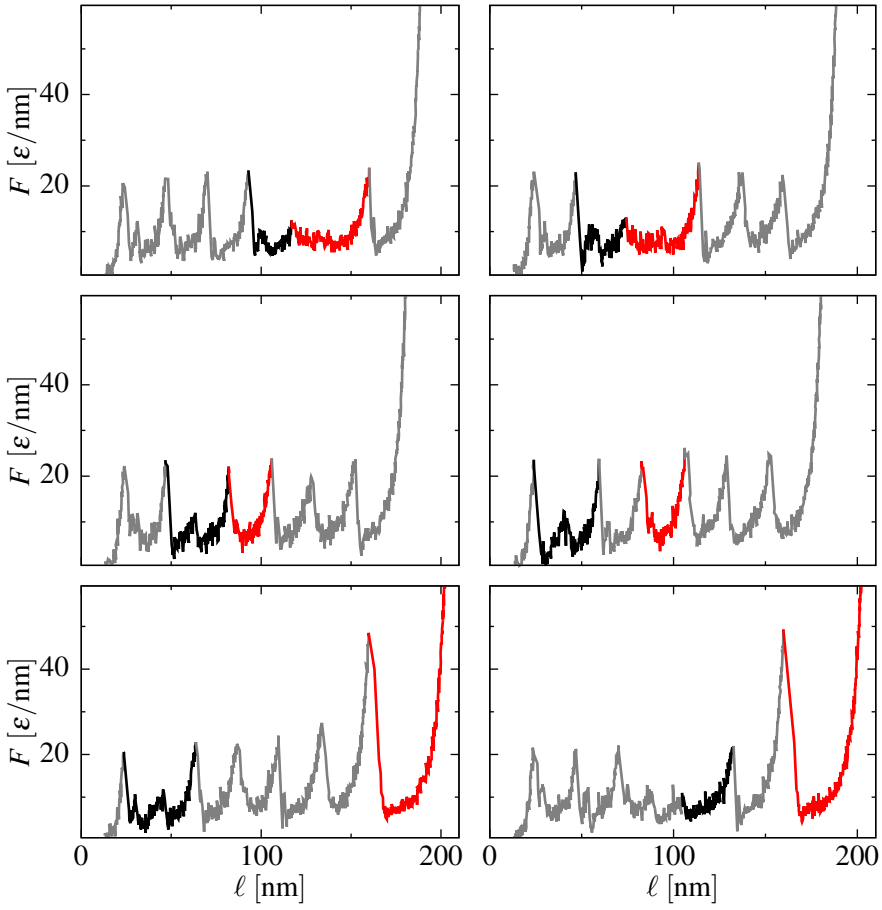


Figure 4.1: Host–Guest studies: *G–H* mechanical stability ratio. Two examples each for the *H–G* strategy of single-molecule markers where the *H* is more (top), similarly (middle) and less (bottom) mechanically stable than the *G*. In each force–extension plot, red marks the unfolded region corresponding to the *G*, black to the *H* and gray to the *M*s.

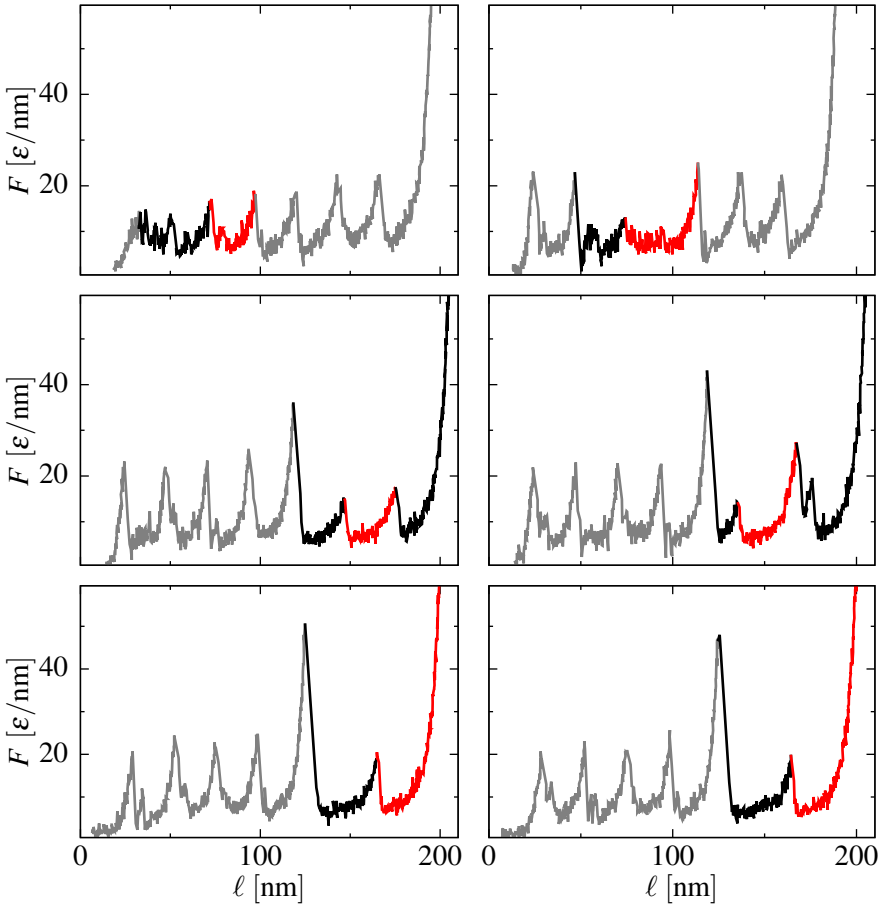


Figure 4.2: Host–Guest studies: *M*–*H* mechanical stability ratio. Two examples each for the *H*–*G* strategy of single-molecule markers where the *H* is less (top), and more (middle and bottom) mechanically stable than the *M*s. The case where the $F_{\max}$ of the *H* and the *M*s are comparable is considered in Fig. 4.1. The middle and bottom panels show a situation where the *H* unfolds with two or one force peaks, respectively. In each force–extension plot, red marks the unfolded region corresponding to the *G*, black to the *H* and gray to the *M*s.

signal is completely contained between the unfolding peak of the H and the end of the curve (which, in the experimental case, corresponds to the detachment peak).

4.3. Summary

In this work, we studied in depth the H–G strategy for single-molecule marking in SMFS. The study showed that the mechanical stability of the G is not affected by it being grafted in the H. It further showed that mechanical protection is needed in order to push the signal of the protein under study away from the noisy proximal region. Moreover, it establishes that the best strategy for clear signalling is to graft the G inside a H with a higher mechanical stability than the Ms used, and that the unfolding of the H should proceed in a single step.

This computational study does not have a direct application in the world of *in silico* SMFS, since single-molecularity and specificity of application points is guaranteed in this case. However, its application to experimental work should increase the efficiency, if not of the data acquisition, of its analysis.

It should be noted that the conclusions of this study yield two constrains on the Ms and the H used in the H–G strategy. Nonetheless, adding further thought to this conclusion, one other condition comes to mind. Even if the computational study simply suggests that the $F_{\max}$ of the H needs to be larger than that of the Ms, these should be chosen wisely. In particular, choosing very sturdy modules as H will increase the probability of having the molecule detach from the cantilever before the unfolding of the H, which would immediately result in a significant decrease of the experimental efficiency. Therefore, the optimal approach is to chose Ms with sufficiently low $F_{\max}$ so that the probability of the Hs unfolding before it is low. An example would be choosing a M with $F_{\max}$ between 20 and 80 pN and a H between 200 and 300 pN.

5. Comparison of the contact maps used in structure-based MD

As stated in Sec. 3.3, one of the key points of structure-based MD is the determination of the contact map. Sometimes the presence or absence of a small group of contacts may have a profound impact on the dynamics, *e.g.* on the folding dynamics of knotted structures [127, 128] or in the determination of the $F_{\max}$ of a protein [129]. Sec. 3.3.1 explains the methodology we use in this work, which has also been used to study virus capsids and protein folding [130], and which we have named overlap (OV). Nonetheless, other methods are also available.

One popular class of methods involves choosing a fixed cutoff length either between the C_α atoms or between pairs of heavy atoms in different residues, and then comparing distances between the atoms to this cutoff. Only those distances that are smaller than the cutoff can generate a contact. More sophisticated variants of this class of contact map generation algorithms involve the effect of shadowing, meaning that the presence of an atom between two others that would be in contact makes that contact disappear [131]. These are quite fast methods, but they are less precise than OV in that the distances between the atoms are not dependent on the atom type.

In this work we focus on the study of a contact map that takes into account the chemical properties of the atoms. This contact map is called Contact for Structural Units (CSU), and we use two versions in this work. The first is the Original CSU (oCSU), as developed in Ref. [132]; while the other one is Repulsion CSU (rCSU), which was designed by us.

5.1. Fundamentals of Contact for Structural Units

The CSU algorithm is carried out in three steps. To begin with, the contacts between all heavy atoms must be found. Thereafter, each atom is assigned a class according to its element and neighbours. Finally, the contacts are classified according to the classes of the atoms involved.

5.1.1. Finding atom–atom contacts

In order to find atom–atom contacts, spheres are assigned to each heavy atom in the protein. Hydrogen atoms are omitted from this step because they cannot be resolved in X-ray crystallography due to their high mobility. The radii of the spheres, R_i, are taken to be equal to the van der Waals radii of the atoms enlarged by the radius of the solvent molecule – 0.14 nm for water. If two of the enlarged spheres overlap – even if the atoms belong to the same residue –, the corresponding atoms are candidates considered for forming a contact. The van der Waals radii used in the CSU server are 0.17 nm, 0.19 nm, 0.15 nm, and 0.19 nm for N, C, O, and S, respectively. Nonetheless, the radii proposed in Ref. [112], which not only depend on the element but also on the bonds it is forming, are more precise than the more general ones in the server. Thus, we have used the more precise values in this work.

In this algorithm, each atom establishes a contact by sharing part of its surface with another atom. In particular, the contact will be established when a solvent molecule cannot fit between the pair. This leads to each atom being able to stablish a discrete number of contacts, typically smaller than the number of atoms that overlapped. To select the real contacting atoms among the candidates, each of the spheres is divided in small uniformly distributed sections using a Fibonacci grid [133]. With this method, the position of the center of section k is given by Eq. 5.1, where F_n is the n'th Fibonacci number and R_i and R_s are the radii of the atom and the solvent, respectively.

$$
\begin{aligned}
r_k &= R_i + R_\mathrm{s} \\
\theta_k &= \arccos\left(1 - 2\frac{k}{\mathrm{F}_n}\right) \\
\phi_k &= 2\ \pi\ k\frac{\mathrm{F}_n}{\mathrm{F}_{n+1}}
\end{aligned}
\tag{5.1}
$$

The index k runs from 1 to F_n. In this work, as in the CSU server, we take $n = 14$, which corresponds to a discretization into F_{14}=610 sections. The corresponding area of each section is 0.0016 of the total area of the sphere.

Once the divisions have been made, overlaps between each subtended volume corresponding to one of the sections and other atoms are sought. If only one atom overlaps with the specific section, the section is assigned that atom and a contact is established. Nonetheless, more than one atom may be overlapping with the same section. In this case, the section is assigned the atom closest to the center of the sphere.

It is important to notice that this criterion breaks the symmetry between the contacts: It may happen that atom i is contacting atom j while atom j is not contacting atom i. An example of such a situation is shown in Fig. 5.1. The lack of symmetry is a violation of Newton's third law of motion and has to be prevented either by removing such contacts or by adding their symmetric partners. In our implementation, we symmetrize oCSU by adding the missing partners and rCSU by removing contacts without their partners. In this way, rCSU is more selective when deciding what is a proper contact. The removal is motivated by the assumption that the surface area corresponding to unbalanced contacts must be small and, therefore, such contacts should be weak.

When all sections have been assigned their corresponding contacting atoms, some of the sections will remain free due to a lack of overlap of their subtended volume with other atoms. These sections correspond to the solvent-accessible area of the atom. Given the residue-based representation of the final contact map, contacts between atoms in the same residue are only useful for computing this area. It should be further noticed that using this method one obtains not only the presence or absence of a contact, but the amount of surface involved in it.

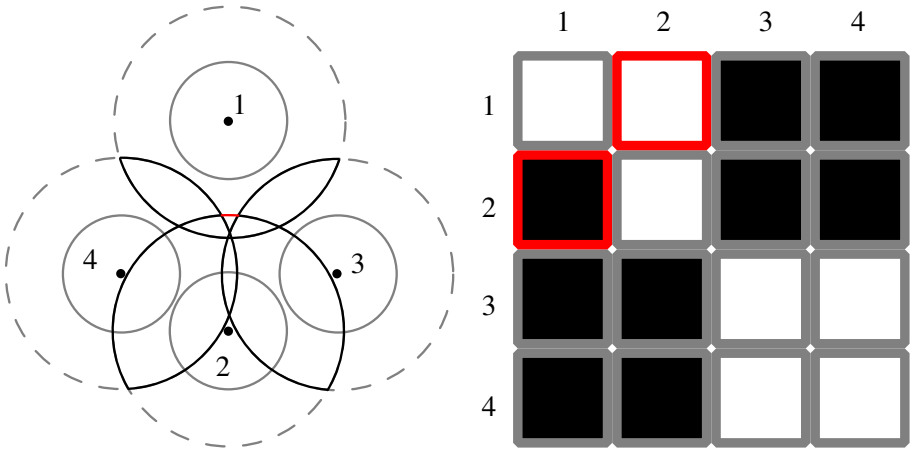


Figure 5.1: Lack of symmetry in CSU contact maps. Left panel: Example of a simple case where CSU yields an asymmetric contact map. The van der Waals volumes of the atoms are painted in gray continuous lines, and the lines corresponding to the enlarged spheres are dashed. Overlaps are highlighted in black and the contact with no symmetric partner is marked in red. Right panel: Graphical representation of the contact map, where contacts are marked in black. The asymmetric contact is marked with a red border.

5.1.2. Assignment of the atom types

The next step in the CSU algorithm is to assign a specific class to each atom. Each class corresponds to a general physicochemical property, such as hydrophobicity or charge. The class depends on the element of the atom as well as its neighbours.

Classes I, II and III correspond to atoms that can be involved in hydrogen bonds. Atoms of class II are hydrogen-bond acceptors and atoms of class III are hydrogen-bond donors. Atoms of class I are those which, due to the lack of further information, must be assumed to be able to act both as donors and acceptors.

Class II is formed by atoms with negative partial charge and no hydrogens attached: The O atoms that have two covalent bonds with other heavy atoms. Class III contains atoms linked to at least one H (or that may have one if the positions of the H's are not provided) and have a sufficiently positive charge to pull most of the electron density from the H, resulting in a situation in which the H has a positive partial charge. All the N atoms, both from the backbone and from the side chain, belong in this class if (and only if) they have less than three covalent bonds to other heavy atoms. Only in this case can N's have a hydrogen attached and act as acceptors. The class-I atoms are all those atoms that present ambiguity in them being acceptors or donors. O atoms that form a bond with just one H belong here, since the H can attach and detach freely in solution. So do the N atoms in the aromatic ring of histidine, which can easily change their protonation state easily.

Hydrophobic atoms belong in class IV: they are not able to create hydrogen bonds nor can they be efficiently solvated by water molecules. All C atoms without bonds to atoms from classes I, II or III and not belonging to an aromatic ring are in this class. The ones that do belong to aromatic rings, independent of their neighbours, are considered to be atoms of class V.

Classes VI, VII and VIII consist of atoms that are neutral. The classification in three groups answers to the fact that their neighbours may balance their neutrality slightly. Thus, C atoms bound to H-bond donors (class II) are neutral-acceptors (class VII) and those bound to H-bond acceptors (class III) are neutral-donors (class VIII). Class VI corresponds to neutral atoms, combining C's bound to class I, C's bound to both class II and class III and also S atoms from cysteine residues.

Next comes one of the differences between oCSU and rCSU. In this work we

decided to differentiate charged residues, which in oCSU are included in classes I, II and III, from the hydrogen-bond-forming ones. This is interesting because electrostatic interaction could in this way be treated differently than hydrogen bonding and yield the location of ionic bridges. In particular, rCSU introduces two new classes of atoms: those which are positively charged (class IX) and negatively charged (class X). Class IX includes the protonated state of the N in histidine as well as the end-of-the-side-chain N in arginine and lysine, while class X is formed by the similarly located O on aspartic acid and glutamic acid.

5.1.3. Contact classification

In the oCSU approach, the atomic contacts are divided into two broad categories: specific or non-specific. The specific ones include hydrogen bonds, aromatic and hydrophobic interactions. Contacts between residues are decided by the presence of at least one specific contact between their individual atoms.

The rCSU algorithm introduces two differences: it includes ionic bridges between atoms of unlike charges, and it recognizes the existence of destabilizing atomic contacts due to the repulsion between the full or partial charges of the same sign. Such a repulsion is considered a non-specific contact in oCSU. Table 5.1 summarizes the contact-assignment algorithm for rCSU. A similar table for the oCSU approach can be obtained by replacing destabilizing contacts (Dc) by the absence of a contact (–) and ionic bridges (Ib) by hydrogen bonds (Hb). In the rCSU algorithm, in order to decide whether a contact between two residues is present, we calculate the number of attractive and repulsive contacts between their respective atoms. If the attractive contacts outnumber the repulsive ones, then a contact between the residues is set.

5.2. Comparison of the different contact maps

Using oCSU and rCSU we obtain different maps than those obtained with OV. Using the information in all of them, we can construct more complete contact maps. Examples of the differences between them are shown for the two proteins most used in SMFS, 1UBQ and 1TIT, in Fig. 5.2 and 5.3 respectively. The different algorithms were used to generate contact maps along 5835 non-redundant protein structures obtained from PDB, and the results show that OV and oCSU yield a similar amount of contacts per residue, (2.0 ± 0.3 and 1.8 ± 0.3), while rCSU gives only 1.05 ± 0.19. Nonetheless, combinations of OV with the two

Class	I	II	III	IV	V	VI	VII	VIII	IX	X
I	Hb	Hb	Hb	Dc	Dc	–	–	–	Hb	Hb
II	Hb	Dc	Hb	Dc	Dc	–	–	–	Hb	Dc
III	Hb	Hb	Dc	Dc	Dc	–	–	–	Dc	Hb
IV	Dc	Dc	Dc	Ph	Ph	–	–	–	Dc	Dc
V	Dc	Dc	Dc	Ph	Ar	–	–	–	Dc	Dc
VI	–	–	–	–	–	–	–	–	–	–
VII	–	–	–	–	–	–	–	–	–	–
VIII	–	–	–	–	–	–	–	–	–	–
IX	Hb	Hb	Dc	Dc	Dc	–	–	–	Dc	Ib
X	Hb	Dc	Hb	Dc	Dc	–	–	–	Ib	Dc

Table 5.1: Classes of interactions in rCSU. A similar table for the oCSU approach can be obtained by replacing destabilizing contacts (Dc) by the absence of a contact and ionic bridges (Ib) by hydrogen bonds.
Atom classes: I – hydrophilic, II – hydrogen bond acceptor, III – hydrogen bond donor, IV – hydrophobic, V – aromatic, VI – neutral, VII – neutral-donor, VIII – neutral-acceptor, IX – positively charged, X – negatively charged.
Types of contacts: Hb – hydrogen-bond, Ph – hydrophobic, Ar – aromatic, Ib – ionic bridge, Dc – destabilizing contact, "–" denotes other (negligible) contacts

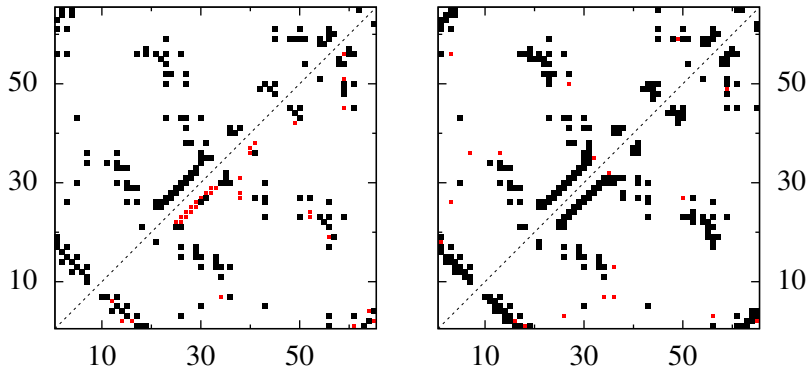


Figure 5.2: Contact maps for 1UBQ. Comparison of different contact maps for 1UBQ. To the left, the two CSU-based methods: oCSU (top) and rCSU (bottom). Black shows the presence of a contact, and red the contacts that are not found with the method but are found with the other version. To the right, OV contact map in black, complemented in red with rCSU (top) and oCSU (bottom).

aforementioned yield OV combined with original CSU (OV+oCSU) and OV combined with repulsion CSU (OV+rCSU), which in turn give rise to 2.3 ± 0.3 and 2.2 ± 0.3, respectively, meaning that both CSU-based methods find contacts that the OV algorithm does not grasp, and *vice versa*.

This fact suggests an obvious question: Which contact map should we use in structure-based MD? We address this question from two perspectives: using folding transitions and stretching simulations. In the former approach, we study the folding of a stretched polypeptide into its native conformation and study the amount of time needed to fold. The most convenient method is the one that provides a wider folding range, as well as a smaller optimal folding time. In the case of stretching, we study the $F_{\max}$ of several proteins that have been studied experimentally, and fit the experimental values with the ones obtained by simulation. The best fit corresponds to a better model.

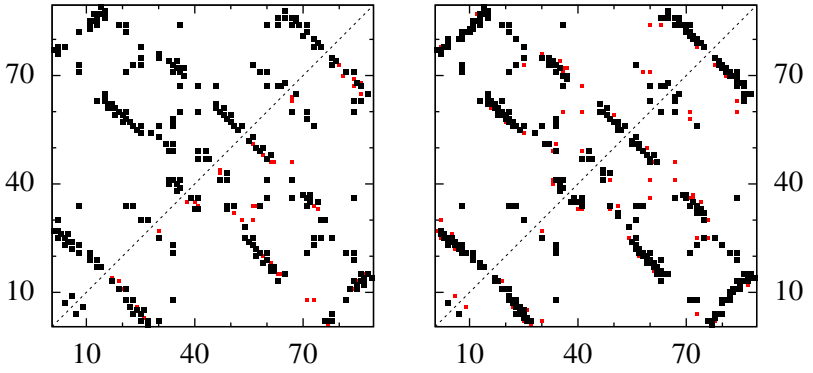


Figure 5.3: Contact maps for 1TIT. Comparison of different contact maps for 1TIT. To the left, the two CSU-based methods: oCSU (top) and rCSU (bottom). Black shows the presence of a contact, and red the contacts that are not found with the method but are found with the other version. To the right, OV contact map in black, complemented in red with rCSU (top) and oCSU (bottom).

5.2.1. Folding studies

The folding simulations are carried out starting from an open conformation, *i.e.* a conformation where none of the contacts are formed, and allow the protein to evolve under the structure-based force field derived from the contact map. The structure is considered folded the first time at which all native contacts are established simultaneously. The median folding time of 300 trajectories for 1TIT and 1UBQ is shown in Fig. 5.4 as a function of T for three contact maps: OV and the combined OV+oCSU and OV+rCSU.

The folding time depends with the temperature as a U-shaped curve, where an optimal temperature – at which the folding time is shortest – is present. The width of the basin of good folding times can be characterized by the two temperatures left and right of the optimal temperature at which the folding time is thrice the optimal value [134]. These temperatures can be readily computed from a fit to a hyperbolic cosine.

From the statistics performed on the folding time of the studied proteins, presented in Tab. 5.2, we can see that the three contact maps are similar in terms of

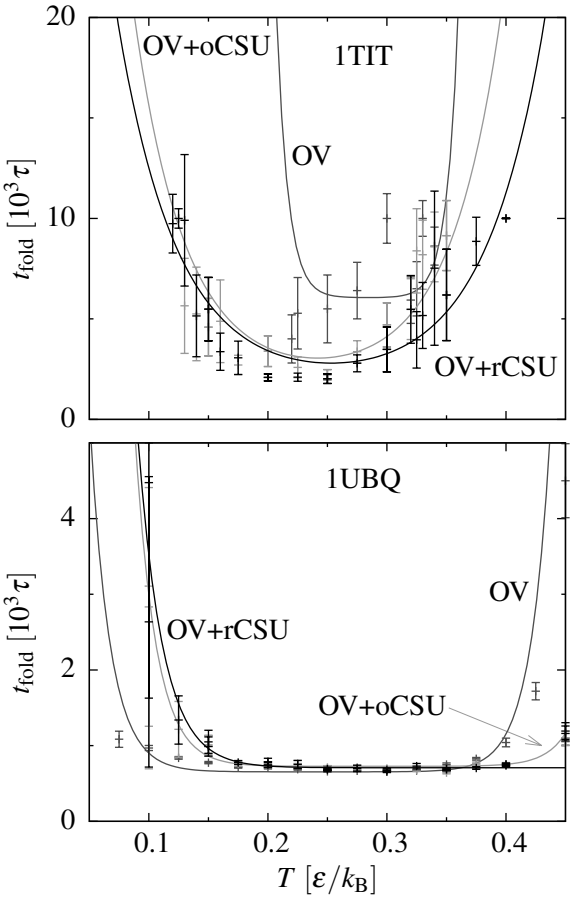


Figure 5.4: Folding times for 1TIT and 1UBQ. Dependence of the folding times for 1TIT and 1UBQ with temperature for different contact maps. Each curve is U-shaped and as such has an optimal folding time.

Contact map	$T_{\min}$ $[\varepsilon/k_B]$	ΔT $[\varepsilon/k_B]$	t_{opt} $[1000\tau]$
OV	0.25±0.01	0.25±0.04	2.66±0.98
OV+oCSU	0.28±0.02	0.24±0.04	2.92±1.27
OV+rCSU	0.27±0.02	0.25±0.04	2.45±1.03

Table 5.2: Folding parameters for different contact maps. $T_{\min}$ is the optimal temperature, at which the folding time is shortest. This folding time corresponds to t_{opt}. ΔT measures the width of the folding curve, measured as the difference between the temperatures at which the folding time is thrice t_{opt}.

width, optimal temperature and optimal folding times. However, for individual proteins like 1TIT, using OV+rCSU or OV+oCSU are clearly a better option, given that even though the optimal times are similar, the folding curve is twice as wide (see Fig. 5.4-top). Thus, it is clear that rCSU adds vital contacts to OV that affect the folding time. This contacts are also found by oCSU, but the latter may also find many other that are redundant. With this we can conclude that rCSU does not improve the performance of the OV algorithm in terms of folding time for a general case, but it is a safer option.

5.2.2. Stretching studies

To study protein unfolding under force, the process is exactly the opposite to the previous folding studies. The N-terminus is fixed at place while the C-terminus is pulled away at a constant speed, as explained in Sec. 3.3. The pulling was performed at $T = 0.3\ \varepsilon/k_B$, given that this is close to room temperature and is always included in the good folding basin. Furthermore, the pulling velocity ranges from 0.01 to 1 pm/τ ($\approx$1 mm/s to $\approx$10 μm/s).

Examples of the curves obtained for 1TIT and 1UBQ pulled at 0.05 pm/τ are shown in Fig. 5.5 for different contact maps. It can be observed that, at least for the highest force peak, the position is not altered while its value is. The variation of the value was expected, given that the number of contacts changed with the contact map. Nonetheless, forces are measured in terms of ε, which needs to be calibrated for each of the contact maps independently. The changes observed in the lower peaks are attributed to the establishment of one or more contacts that bridge the gap between the peaks.

In order to calibrate the value of ε for each of the models, we studied 38

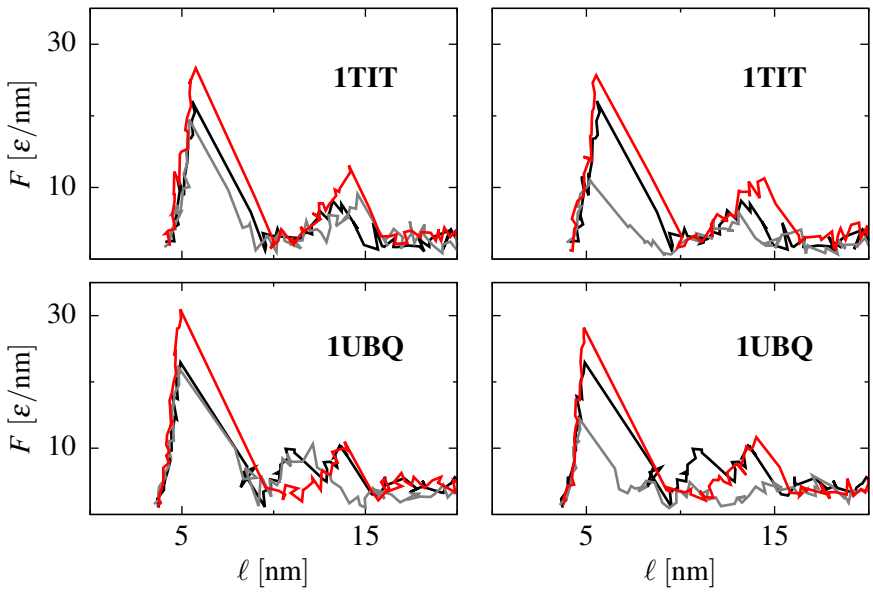


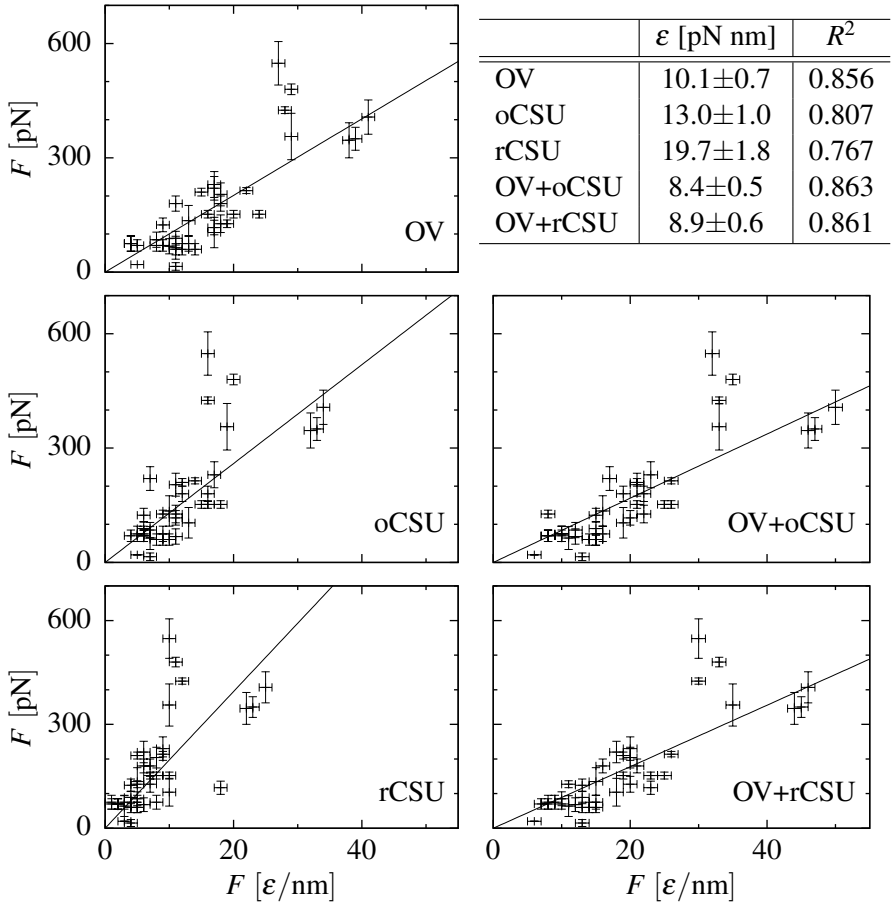
Figure 5.5: Unfolding curves for different contact maps. The unfolded proteins are 1TIT (top) and 1UBQ (bottom). The left panels show the differences between OV (black) to oCSU (gray) and OV+oCSU (red).The right panels compare OV (black), rCSU (gray) and OV+rCSU (red). The pullings were done at a speed of 0.05 pm/τ. It can be seen how the combined contact maps (red) yield higher force peaks and the CSU-based ones (gray) yield them smaller – especially rCSU –, but the position of the resistant elements in the protein do not depend on the contact map.

proteins that had been studied experimentally and compared our results with the ones from the experiments. To do that, $F_{\max}$ needed to be rescaled according to the pulling speed of the experiments using a logarithmic law [77]. After rescaling each of the values, we fitted a zero-intercept line to each of the cases and obtained the values of the slope as the conversion factor between pN-measured experimental forces and theoretical forces measured in ε/nm. The corresponding scatter plots for all of the contact maps studied and the table summarizing the values of ε can be found in Fig. 5.6. Looking at the goodness of the fits – assessed by R^2 –, the combined models perform the best in unfolding, closely followed by OV.

5.3. Summary

In this work, we introduced a new way of studying the contact map of proteins based on the CSU approach, which we called Repulsion CSU (rCSU) due to the fact that it takes into account the repulsion between charges, when the original method does not. rCSU selects the well defined contacts, and even if it does not correlate well with the experimental data on protein stretching, in combination with the OV contact map it works better than OV alone. This is probably due to the addition of a few important contacts that the OV algorithm misses altogether.

The contact map algorithm developed here should, however, be developed further in order to be complete. For example, the fact that one atom has many sections involved in a contact could probably be taken into consideration in the ε parameter, as could as well the type of contact that two residues are establishing (*e.g.* ionic bridges could be stronger than hydrogen bonds). Nonetheless, for the time being, it might be safe to use the combined OV+rCSU algorithm for contact map generation in structure-based molecular dynamics studies of protein folding/unfolding.



	ε [pN nm]	R^2
OV	10.1±0.7	0.856
oCSU	13.0±1.0	0.807
rCSU	19.7±1.8	0.767
OV+oCSU	8.4±0.5	0.863
OV+rCSU	8.9±0.6	0.861

Figure 5.6: Calibration of ε. Each scatter plot shows the experimentally measured force (ordinates) *vs.* the theoretically computed one (abscissas), once it has been rescaled to the corresponding pulling speed. The table at the top right panel shows the values of the slopes of each of the fits together with the statistical error and the R^2 Pearson coefficient.

6. The conformational polymorphism of neurotoxic proteins

Amyloids, including the toxic ones, have been demonstrated to share a fair amount of common traits along their aggregation cascade. Specifically, they show similar fibers as described by the cross-β spine structure in X-ray crystallography [25] and similar oligomeric structures as recognized by the same conformational antibody, A11 [26]. However, the study of the common traits at the monomeric level has been elusive to high-resolution techniques, since the fluctuation of each of the monomeric species is too fast and thus too many conformations coexist in the population, although they have been able to report a high degree of α-helical and random-coiled folds [135, 136, 137].

As mentioned in Sec. 1.2.3, single-molecule techniques such as AFM can give insight on the population of a sample rather than an average. Therefore, we performed a single-molecule AFM study of the proteins causally related to specific diseases such as Huntington and Alzheimer.

Previous such studies had been performed, but lack the necessary controls to study such proteins [138, 139, 140]. In particular, the controls they missed include the assertion of the intramolecular nature of the interactions, the differentiation of the signals from noisy proximal events and the corroboration of

	Q_{19}	Q_{35}	Q_{62}	Q_{62}^{+}	Aβ	DM	Arc	Arc^{+}
N	111	100	107	124	244	396	102	108
$F_{max}>20$ pN [%]	0	5$\pm$3	7$\pm$4	3$\pm$2	20$\pm$10	0	32$\pm$16	30$\pm$15
max(F_{max}) [pN]	<20	420	720	200	530	0	530	310

Table 6.1: Experimental results on SMFS of IDPs. The superindex + marks the presence of QBP1, DM stands for Double Mutant, which corresponds to F19S/L34P Aβ, and Arc is the arctic mutation E22G Aβ.

the actual amyloidogenic ability of the studied constructs. We have avoided all these limitations by using the *H-G* strategy as explained in Sec. 2.4.2 along with stringent criteria for data selection that ensure that the data are reliable [31].

6.1. Mechanical polymorphism of polyglutamine

We start these studies using Q_n, a stretch of glutamines that is responsible for several diseases known as polyglutaminopathies, including Huntington and several Spinocerebellar Ataxias, all of which develop only once a threshold in the length of the polyglutamine chain has been surpassed. For most of these diseases, the threshold is around 35 repeats. We studied Q_{19}, Q_{35} and Q_{62}; thus below, close to and above the threshold, to examine the possible differences in the monomeric species depending on their disease-inducing capabilities.

After pulling the different constructs in the AFM, the highest force peak corresponding to each of the molecules studied was registered, and the results are summarized in Fig. 6.1 and Tab. 6.1. Q_{19} presented no molecules with a detectable force peak, while Q_{35} and Q_{62} presented 5$\pm$3 % and 7$\pm$4 % of the molecules with at least one force peak significantly above the detection limit of our AFM (20 pN).

With this analysis one observes that the presence of mechanically stable conformers is directly related to n. Nonetheless, the specific relation cannot be established due to the lack of sufficient n values. Furthermore, the maximum F_{max} observed in each case also correlates with the length, being undetectable (<20 pN) for Q_{19} but 420 pN and 720 pN for Q_{35} and Q_{62}, respectively.

In an attempt to assess the relationship of the mechanostable conformers to the disease, we used a peptide specifically designed to attach to Q_n chains, Glutamine Binding Peptide 1 (QBP1), which has been proved to inhibit amyloido-

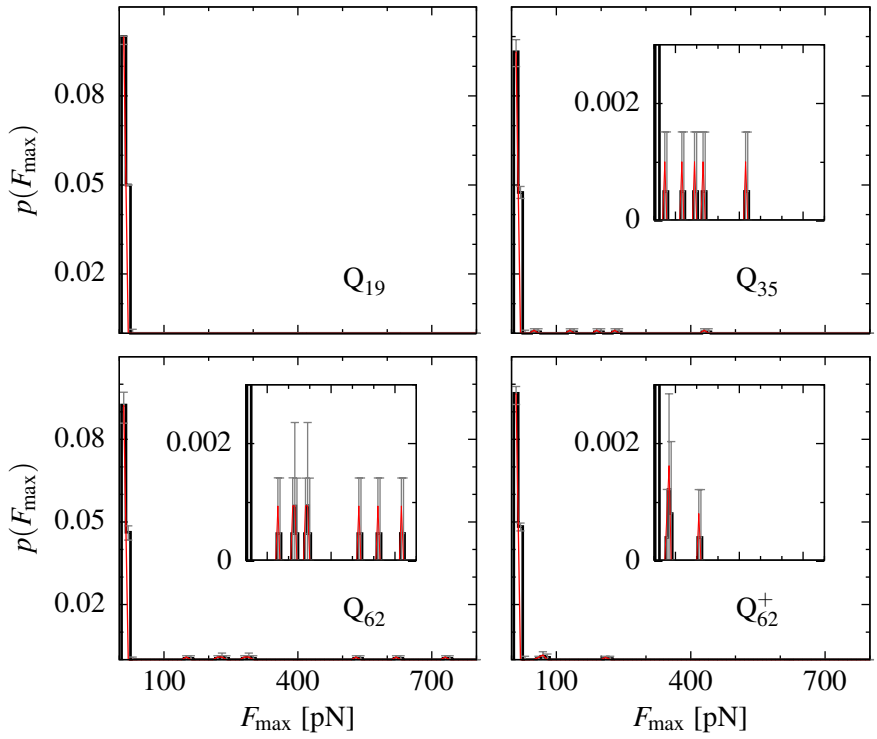


Figure 6.1: Experimental F_{max} histogram of polyglutamine. Each graph shows the histogram as obtained experimentally in red, and a 95 % confidence interval in black). The insets show a zoom-in on the low-probability events. The case of Q_{19} has no inset because there are no events with mechanical stability higher than the noise. Q_{62}^{+} stands for Q_{62} with the inhibitor QBP1.

genesis and to revert neurodegeneration caused by expanded Q_n in *Drosophila melanogaster* [24]. The incubation of Q_{62} with QBP1 reduced the number of mechanically stable conformers found (mechanical polymorphism) down to 3 ± 2 % and the maximum $F_{\max}$ to 200 pN. Thus, both the maximum $F_{\max}$ and the number of mechanostable conformers appear to be importantly related to the development of a Q_n-related disease.

6.2. Mechanical polymorphism of β-amyloid

Similar to Q_n, Aβ was studied using SMFS. In this case, no correlation with length is available, but we added mutations to the study. In particular, we studied the wild type Aβ42, known to be more toxic than its 40-residue homologue, and two mutations on it: F19S/L34P, a non-fibrillogenic one; and E22G –also known as arctic mutation–, a familiar mutation that produces earlier onset and more aggressive Alzheimer. The QBP1 inhibitor, which we found to also affect the aggregation in other IDPs [31], was used on the arctic mutant so as to see its effect on the polymorphism of Aβ. The results are summarized in Tab. 6.1 and Fig. 6.2.

Interestingly enough, the polymorphism presented in Aβ is much greater than that seen in Q_n, which reinforces the results of the latter. On the other hand, QBP1 has no effect when applied on the arctic mutation, since it does not change the shape of the distribution significantly even if it removes the lone high-force event. It is therefore clear that more aggressive familiar mutations correlate with a greater polymorphism; while a non-fibrillating mutant, similar to the innocuous Q_{19}, presents no polymorphism at all.

6.3. Summary

Taking the results from Aβ and Q_n together, we can see that the common characteristics present among different amyloids is not only from the oligomeric stage onwards but starts with a rich conformational polymorphism in the monomer. We further prove that the polymorphism is exacerbated in proteins involved in more severe cases of the disease, such as pathogenic Q_n chains and familiar mutations like E22G Aβ, and is abolished when the proteins are not disease-inducing or not fibrillogenic, such as Q_{19} and F19S/L34P Aβ.

Furthermore, the fact that an inhibitor is capable of both reducing the poly-

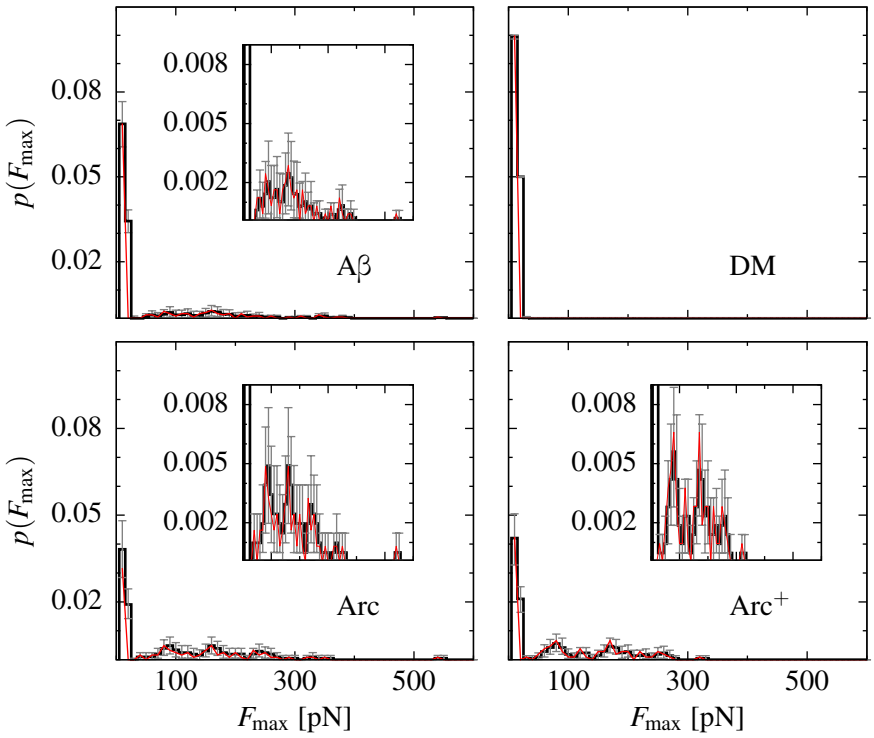


Figure 6.2: Experimental F_{max} histogram of β-amyloid. Similar to Fig. 6.1, the red line shows the experimental results and the black bars a 95 % confidence interval. DM represents the F19S/L34P mutation and Arc stands for arctic (E22G). The latter has been also tested with QBP1 (Arc$^+$).

morphism and stopping fibrillation supports the conformational change hypothesis [141], whereby the monomer would undergo a conformational change that would trigger the oligomerization. This conformational change would be impeded by QBP1 at least in the case of Q_n. The fact that QBP1 has no effect on Aβ might be related to the fact that Aβ is too short when compared to Q_{62}, but may also indicate that Aβ can undergo a different conformational change, unaffected by QBP1, which leads to fibrillation.

7. The universe of conformers of neurotoxic proteins

It is clear from the experimental results that the events that present high forces are not frequent. This is especially so in the case of Q_n, where we have been able to find up to 7 % (in Q_{62}) with detectable force. However, the structures adopted by the mechanostable conformers remain elusive to us, so we tackle the issue using BEMD (see Sec. 3.2).

7.1. Generation and selection of the independent conformers

The BEMD simulation generated a set of six trajectories (one for each replica), from which we extracted one snapshot every 5 ps. As aforementioned, a snapshot consists of the positions and velocities of all the particles in the system at that time. Once the snapshots were selected, a three-sieve method was applied in order to obtain structures that were temporally and structurally independent: In the first step, the DSSP algorithm [142] was used to obtain the Secondary Structure Content ($\mathbb{SS}$) for each snapshot. $\mathbb{SS}$ is defined as the number of residues in the protein that belong to an α-helical region, a β sheet or a hydrogen-bonded

turn, normalized to the chain length. Since the conformers we are interested in are those that might yield force in a SMFS, the first sieve eliminated those snapshots with less than 30 % $\mathbb{SS}$, while the remaining ones were forwarded to the second sieve.

The second sieve was used to find temporally-uncorrelated structures from those obtained in the first stage. To this end, we chose the division of a time cluster to be at the point where two structured conformers were separated by at least 50 ps of unstructured ones. The conformer in each cluster with higher $\mathbb{SS}$ was chosen to represent the cluster and proceeded to the third and final sieve. Fig. 7.1 presents an example of the first and second sieve of one of the replicas.

The third sieve checked for structural independence, and was carried out on all time-cluster representatives at the same time irrespective of the replica they originated from. In order to study this feature, we used the TM-score [143] and our version of the TM-align algorithm [144] in which the determination of the secondary structure is based on the results originated from DSSP, which has been shown to perform better [123].

The TM-score value measures similarity between two proteins according to an atomic alignment, which can be based either on sequence or secondary structure. In this case, sequential alignment is not needed since all conformers contain the same sequence. After the alignment is done, only the aligned atoms are taken into account by summing the inverse of the distances between them, then normalizing to the total length of the protein, as in Eq. 7.1, where n_a is the number of aligned atoms, n is the length of the protein, d_i is the distance between the atoms in the i-th pair and d_0 is a standard distance used for normalization. The max function refers to all possible alignments.

$$TM = \max \left[\frac{1}{n} \sum_{n_a} \frac{1}{1 + \left(\frac{d_i}{d_0} \right)^2} \right] \tag{7.1}$$

Therefore, the TM-score would be 100 % if all the atoms in the proteins could be aligned with one another, and in this alignment they all were at the same position (meaning the distance between them would be 0). This process results in a matrix of scores that rate their similarity in a pair-wise fashion. As in Ref. [123], two conformers are considered neighbors if their TM-score is greater than 45 %.

At this stage, we identified the conformation which has the highest number of

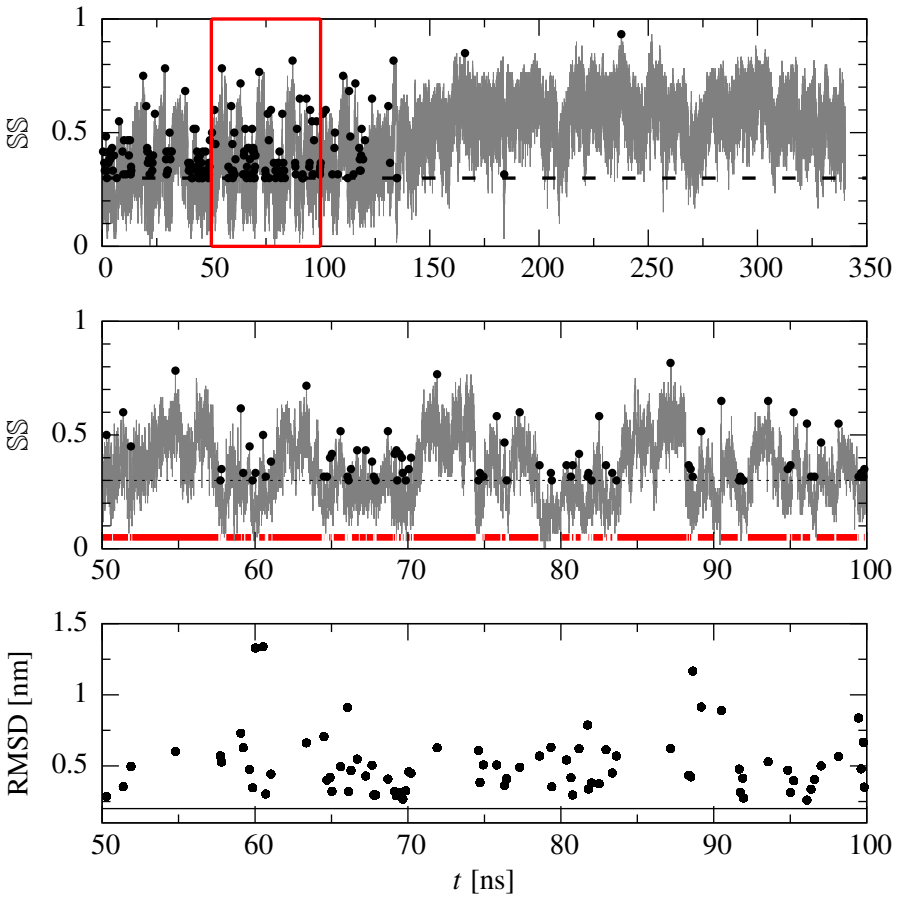


Figure 7.1: Example of the first and second sieving stages. The gray line in the top panel shows the evolution of $\mathbb{SS}$ with time for one of the replicas. Structures with $\mathbb{SS} > 30$ % (the thin horizontal line) are taken for clustering. A cluster ends whenever the gap between successive structured conformers becomes greater than 50 ps. The black dots correspond to the structures with highest $\mathbb{SS}$ in each cluster, which are chosen as representatives. The red box in the top panel is shown zoomed in the middle panel, where clusters are represented by red lines. The bottom panel shows the RMSD of each cluster representative relative to the previous one. All of these RMSDs are greater than 0.2 nm so the clusters can be considered to be uncorrelated in time [102].

	Q_{16}	Q_{20}	Q_{25}	Q_{30}	Q_{33}	Q_{38}	Q_{40}	Q_{60}	Q_{80}
t [μs]	0.38	0.66	0.40	0.84	1.19	0.91	0.83	2.04	1.76
N	298	491	330	422	479	322	269	246	108
$\min(\langle z \rangle)$	5.25	5.4	5.52	5.60	5.64	5.68	5.7	5.8	5.85

Table 7.1: Characteristics of the Q_n independent conformers. For each of the chain lengths studied in the case of Q_n, total simulation time, number of independent conformers found and simply stiff limit.

neighbors. This conformation and its neighbors were denoted as a cluster, which is removed from the pool. The conformer in the cluster with the largest $\mathbb{SS}$ was selected to belong to the final set of structures. This procedure was then repeated with the conformations still remaining in the pool, until the pool became empty. Tab. 7.1 shows the amount of conformers obtained for each of the chain lengths studied.

After clustering, all independent structures underwent a steepest descent minimization process until the maximum force between a pair of atoms was smaller than 0.25 J/(mol nm) so that the structure in the closest energy minimum was obtained. In this process, some of the residues may form or break contacts, thus changing their secondary structure content slightly. Therefore, even though the structures were selected with $\mathbb{SS} \geq 30$ % before the first clustering, some of the final structures may have a smaller $\mathbb{SS}$ content.

7.2. Conformer descriptors

In order to characterize each conformer we used several structural and dynamical descriptors. Among the structural description, we used geometrical ones such as the radius of gyration, R_g, which characterizes the linear size, and the w parameter, which describes the shape [145, 146]. The former, as demonstrated in Eq. 7.2, is computed as the average of the radius of the spheres that would enclose all the atoms in the system respect to its geometrical center. The w parameter is in turn defined through the diagonalization of the tensor of inertia and by making combinations of the three main radii (R_i), as exposed in Eq. 7.3. This parameter is such that a near-zero w corresponds to a globular shape, a positive w to an elongated conformation, and a negative w to a flattened object.

$$R_g = \frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \langle\vec{r}\rangle) \tag{7.2}$$

$$w = \frac{2R_2}{R_1 + R_3} - 1;\ i < j \Leftrightarrow R_i < R_j \tag{7.3}$$

Next we characterized the secondary structure of the protein using the DSSP procedure [142], and obtained the percentage of residues belonging to a structured region as the $\mathbb{SS}$ parameter.

Finally we used two other descriptors: $F_{\max}$ and $\langle z\rangle$ – the average coordination number. The former relates to the dynamics directly, while the latter relates to it indirectly since z measures the number of residues a given residue interacts with, interaction being either through the peptide bond with its two nearest neighbors along the sequence or through contact interactions with residues which are further away in the sequence. The contacts play a dynamical role in coarse-grained structure-based models but they can also be used as descriptors in all-atom models. The specific definition of the contacts we use is based on the OV algorithm (see Sec. 3.3.1).

Maxwell demonstrated [147] that the temporal stability of three-dimensional systems of particles with pairwise interactions depends on the $\langle z\rangle$. Following the terminology used by Maxwell, if we define an n-particle system with pairwise interactions to be *stiff* if it contains no pairs of particles that can be moved apart without affecting a bond; and a stiff system is considered *simply stiff* if the removal of any bond will turn it into being not stiff, then the stiffness of the n-particle system depends on the dimensionality (D) of space in which the particles are set and on the number of bonds (b) between them. Instead of b, we can discuss the dependence on the average coordination number because the two quantities are closely related. It is easy to see that the sum of the coordination numbers of all of the particles in the system is equal to the double of the number of bonds. This is because each bond connects two particles and thus it counts twice. Thus, the average coordination number for the system is given by Eq. 7.4.

$$\langle z\rangle = \frac{2b}{n} \tag{7.4}$$

For particles moving along a line (one-dimensional system, one degree of freedom), a system is simply stiff if $b = n - 1$. This equation can be proved by

using the method of mathematical induction. For $n = 2$, the system is simply stiff when one bond is present. If we have a simply stiff system of n particles and we add one more, then only one extra bond is needed to ensure that this new particle will not be able to move away from any other. Therefore, we can derive Eq. 7.5, the average coordination number for a simply stiff 1D system, from Eq. 7.4.

$$\langle z \rangle_{1D} = 2 - \frac{2}{n} \tag{7.5}$$

For 2D and 3D systems, the number of bonds in a simply stiff particle system is $b = 2n - 3$ and $b = 3n - 6$, respectively. The proof can be obtained in analogy to the 1D case. Therefore, the average threshold coordination number is given by Eq. 7.6 and 7.7.

$$\langle z \rangle_{2D} = 4 - \frac{6}{n} \tag{7.6}$$

$$\langle z \rangle_{3D} = 6 - \frac{12}{n} \tag{7.7}$$

In the thermodynamic limit ($n \to \infty$) of a 3D system the threshold $\langle z \rangle$ is 6, as shown by Maxwell. For finite protein-like systems, this threshold value is reduced. In the cases of this study, simply stiff limits are listed in Tab. 7.1.

As for F_{max}, it is determined as the highest force peak in a force–displacement curve. It is possible, however, that no articulated force peak appears in a trace – meaning it does not exceed the thermal noise level of about 1 ε/nm – before the force raises indefinitely due to stretching of the peptide bonds. In this case, similar to the experiments, the force was assumed to be zero and the conformer was considered non-mechanostable.

Even if $\langle z \rangle$ and F_{max} seem intuitively related, cases might occur where low $\langle z \rangle$ generate high forces – *e.g.* in a case where few local contacts generate a mechanical clamp – and *vice versa* – *e.g.* long α helices that present many contacts but are easy to unfold mechanically.

7.3. Structural and dynamical analysis of Q_n

We first considered Q_{20} and Q_{60}, so that one can compare with the experimental results on Q_{19} and Q_{62} in Sec. 6. To better put them in context, we also studied Q_{40} and Q_{80}. Fig. 7.2 represents, for these sets, the geometries obtained on the

	Q_{20}	Q_{40}	Q_{60}	Q_{80}
$F_{\max} > 0$ [%]	24±3	51±3	61±3	75±2
$\max(F_{\max})$ [ε/nm]	16	21	25	29
$\max(F_{\max})$ [pN]	176	231	275	319
$\langle z \rangle < \min(\langle z \rangle)$ [%]	43.0±1.0	16±2	12±2	12.5±1.7

Table 7.2: Dynamical parameters of Q_n. The table shows the number of mechanostable conformers, the maximum $F_{\max}$ and the number of volatile conformers (assessed by $\langle z \rangle$ being smaller than the stiff limit) for each studied length.

$R_g - w$ plane, as well as scatter plots of $F_{\max}$ *vs.* $\langle z \rangle$ and $\mathbb{SS}$. The graphs show how the shape of the molecules is similar even if their size grows with n, and further shows that the forces are unrelated to either $\langle z \rangle$ or $\mathbb{SS}$. In particular, some of the top five structures regarding $F_{\max}$ have low $\langle z \rangle$ (even lower than the simply stiff limit in the case of Q_{60}) and low $\mathbb{SS}$. This is because typical high-force motifs include β-structured regions [148], while high $\langle z \rangle$ and $\mathbb{SS}$ can be achieved with α-structure and hydrogen-bonded turns.

Furthermore, in Fig. 7.3 there is a comparison of the distributions of $F_{\max}$. We observe that even though BEMD simulations bias the chain towards the acquisition of $\mathbb{SS}$, many conformers do not produce any articulated force peaks above the noise level. Tab. 7.2 collects the amount of mechanostable conformers from each set. It should be noted that the experimental results in Sec. 6 yield different $F_{\max}$ distributions due to the intrinsic nature of the BEMD simulations and the sieving protocol, which selects only the most structured conformers and is thus prone to select mechanically resistant ones. Moreover, the fact that the theoretical values are much smaller than the ones found experimentally can be attributed to a small statistics, since the experimental systems yielded high force only with extremely low probability ($p(F_{\max} > 200 \text{ pN}) = 7 \pm 6$ %). This results are thus consistent with the experimental data, where no force peaks were detected in Q_{19}, while some were found in Q_{62}. Remarkably, although the diversity in mechanical stability grows with n, the frequency of independent structure generation has an opposite relation, *i.e.* smaller chains yield more independent conformers in less time, as can be seen in Tab. 7.1, and so does the conformational polymorphism. The volatility of each conformer assessed by $\langle z \rangle$ lower than their threshold, reflected in Tab. 7.2, also agrees with this conclusion.

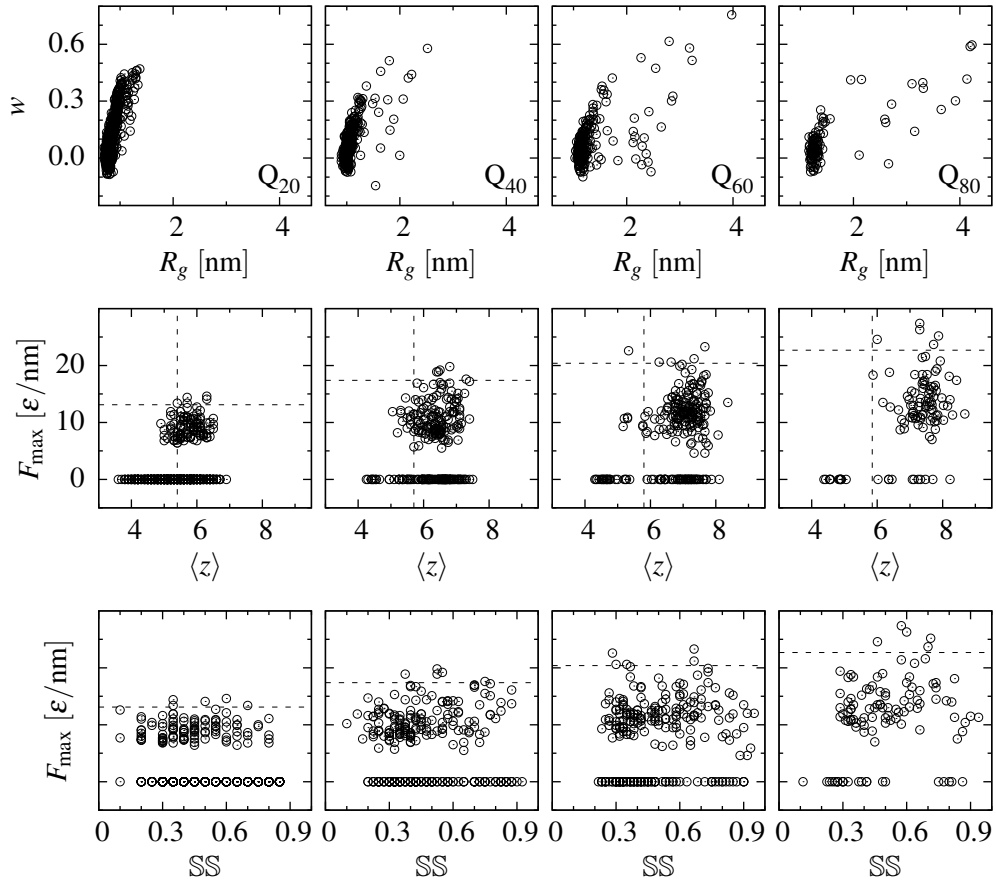


Figure 7.2: Structural and dynamical characterization of Q_n. The top row shows the dispersion of conformers in the $R_g - w$ plane for the representative Q_n sets: Q_{20}, Q_{40}, Q_{60} and Q_{80}. The second and third rows show the dispersion of F_{max} *vs.* the average coordination number and Secondary Structure Content, respectively. The horizontal lines mark the top five conformers as classified by F_{max}, while the vertical lines mark the simply stiff limit for each set.

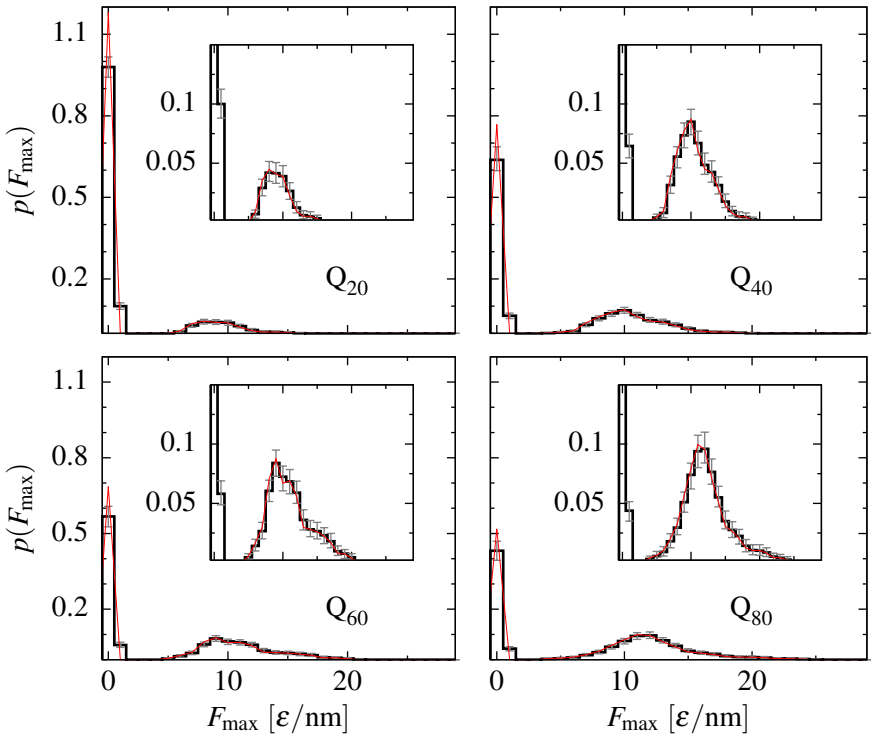


Figure 7.3: Distributions of $F_{\max}$ for Q_n from simulations. The red lines show the obtained distribution, while the black lines show a 95 % confidence interval. The insets show a close-up look at the distribution of forces for the mechanostable conformers.

Taken together, these results show that $F_{\max}$ is inherently different in the different Q_n's, even when they are structurally similar. This further points to $F_{\max}$ not being related to either to $\mathbb{SS}$ or $\langle z \rangle$.

7.3.1. Life span of the structures

In order to test the rellevance of these conformations, we performed 10 ns free-dynamics simulations on 100 structures chosen randomly from each Q_n set. We studied the time dependence of RMSD relative to the initial structure and the last time that it fluctuated below 0.2 nm was recorded for each conformer as its time of residence (t_R). Similarly, we define the escape probability ($P_e(t)$) as the probability of leaving the initial conformation before time t. Fig. 7.4 shows the results of this study: increasing n makes Q_n conformers last longer in a specific state, with significant differences between the different studied groups of $n = 20$, 40, 60 and 80.

Interestingly, both theoretical and especially experimental pulling experiments are typically done at pulling velocities such that the time the protein is being pulled is far longer than 10 ns. In particular, the pulling simulations performed in this work take ≈ 50 μs to completely extend a protein with 60 residues, while the experiments take around 60 ms to accomplish the same task. This leads to question whether the force peaks present in the experimental traces really relate to the initial conformers or to structures that have actually been formed while the molecule was being pulled. Therefore, one must look at $F_{\max}$ carefully since it has different meaning in this kind of simulation than in the experiments: Here, $F_{\max}$ is associated directly with a conformer, since simulations are based on the initial contact map. On the other hand, in experiments, molecules are subjected to fluctuations with a characteristic time of 1 ns and the force–distance curves carry information not only about the initial conformer but also about the stretching-unrelated intrinsic shape transformations that the protein may undergo.

All in all, we observe that disordered proteins such as Q_n's are not long lasting, and that mechanical stabilities need to be looked at in the context of how they were measured, either referred to the initial conformer if done through structure-based modelling, or including bond formation during the stretching if performed experimentally.

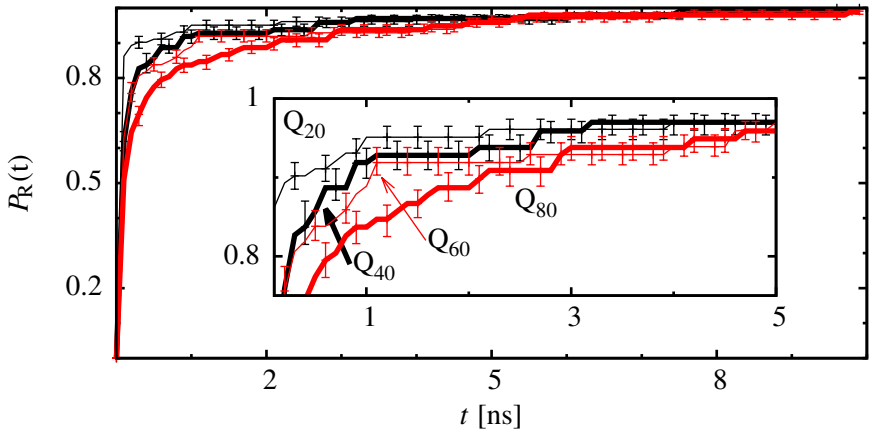


Figure 7.4: Time evolution of Q_n. For each n, 100 randomly chosen structures have been placed under a free-dynamics evolution for 10 ns. After that, the RMSD has been studied and the last time when it fluctuates above .2 nm is recorded as the residence time . The graph shows the escape probability ($P_e(t)$), defined as the probability of the residence time being smaller than t. We can see how n governs the escape probability for Q_n, making Q_{20} fluctuates out of the initial structure faster than Q_{40}, which is in turn faster than Q_{60} and Q_{80} being the fastest.

7.3.2. Structures with knots

One of the criteria when generating structures to start the BEMD simulations was that the random model was not knotted. Nonetheless, some of the independent conformers obtained from them did present knots.

Knots were only present in the Q_{60} set and with a low probability (9.3 ± 1.8 %). Furthermore, two types of knotting were found, the most abundant being trefoil (3_1), and the other less populated the three-twist (5_2). Examples of these knots can be found in Fig. 7.5. Upon stretching, only (13 ± 7) % of the Q_{60} knotted structures untied. As shown in Ref. [149], tightening of knots may be associated with force peaks which in the case of Q_{60} had heights from 9 to 24 ε/nm. Importantly, knotted structures would have been found experimentally but not included in the analysis, since the final length would be shorter than expected. This kind of reccordings were classified in Ref. [31] as putative mechanostable events.

Fig. 7.5 further shows a histogram of the ends of the knots (k_-, k_+) and their extension (Δk, measured as the number of residues contained inside the knot) in the conformers. It is interesting to note that the average extension of the knotted Q_{60} conformers is 36 (with a 0.12 % error), which corresponds to the median threshold value for most polyglutamine-expansion-related diseases including Huntington. Moreover, even if knotted proteins found in nature are normally in enzymes (transferases, anhydrases, synthetases...), the only hypothesized function of the knot itself – as opposed to the whole protein – is for human ubiquitin hydrolase UCH-L3, a protein in charge of the de-ubiquitination of the proteins labelled to be degraded in the proteasome. The function that has been suggested for the knot in this case is to prevent the unfolding of the protein in a case where the proteasome were to try and degrade it [150].

This observations lead to think of the knotted conformations as one of the pathways to toxicity of Q_n: The knots might jam the proteasome and prevent from other misfolded proteins to be degraded for a long time. However, in order for this to be so one would expect to also find knots in the other studied toxic species (Q_{40} and Q_{80}). The fact that they were not found can be attributed to a low probability of knot formation combined with small statistics, which would imply that BEMD took Q_{60} through a knot-forming path while taking the rest of Q_n studied through non-forming ones. Therefore, an increase in the sampling may catch these knotted structures in Q_{80} and Q_{40}, while their formation is fairly improbable for n below 35 since the typical knot size is about this length.

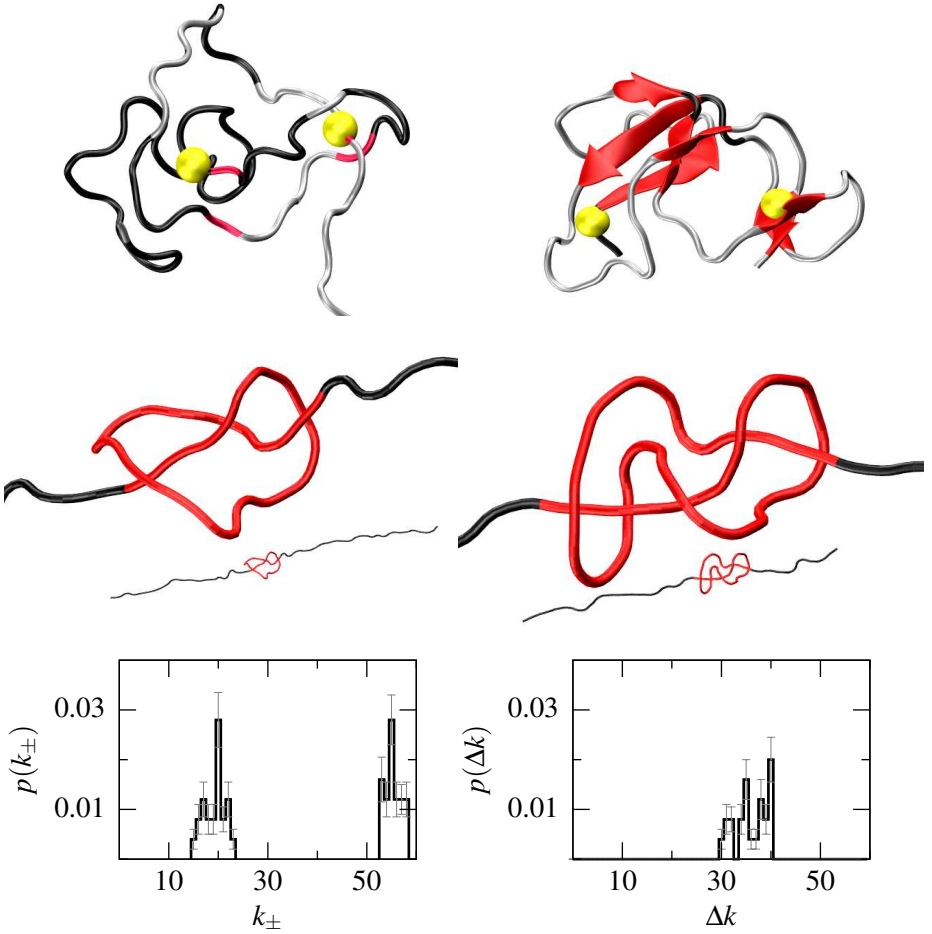


Figure 7.5: Knots in Q_{60}. The top panels show examples of a trefoil (3_1, left) and a three-twist (5_2, right) knots with the knot ends highlighted with yellow spheres. The middle panels present the same conformations having been partially stretched, with the region inside the knot highlighted in red and zoomed in. The bottom panels represent histograms of the knot end positions ($k_\pm$) and their corresponding extension (Δk).

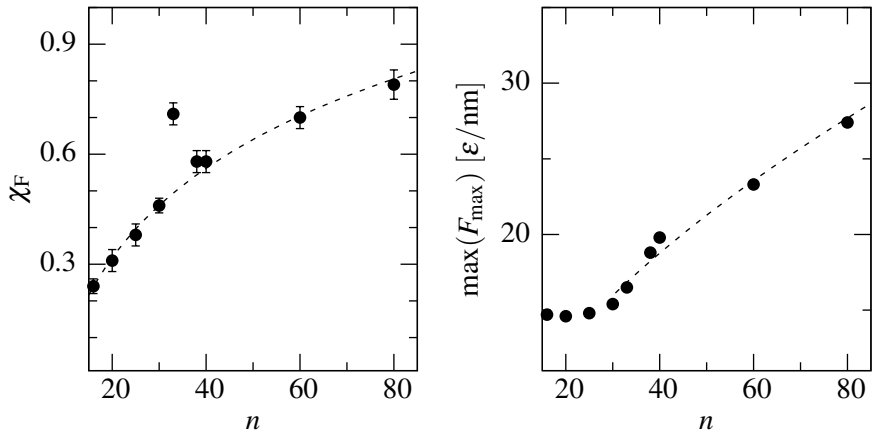


Figure 7.6: Mechanical stability of Q_n as a function of n. χ_F represents the fraction of conformers with at least one force peak for that particular length. The dotted fits correspond to a logarithmic function (left) and a polynomial behavior (right), the latter being typical for avalanches.

7.3.3. Q_n in a wider context

In order to further scrutinize the differences at the disease thresholds, new sets of Q_n were generated with $n = 16$, 25, 33 and 38. Fig. 7.6 shows the evolution of the mechanical stability. In particular, the fraction of conformers with $F_{\max} > 0$, which we name χ_F, follows a logarithmic law, while the maximum $F_{\max}$ for each set behaves like an avalanche system: it has a constant value until $n = 33$, and then starts growing as a power law with exponent 0.562.

Furthermore, taking advantage of the previous work on polyvaline [123], and using the CATH database [107], we compared the features of Q_n to other groups of proteins. In particular, given that the polyvaline studied in [123] is of length 60, we compared the sets of Q_{60}, V_{60} and $CATH_{60}$, the latter defined as those proteins from the CATH database with 57 to 63 residues. The V_{60} from Ref. [123] came from a 50 μs simulation and after clustering yielded 7076 independent structures. $CATH_{60}$ was downloaded from the CATH database and contains 256 proteins.

Fig. 7.7 shows a scatter plot of R_g *vs.* w and a $F_{\max}$ histogram for V_{60} and $CATH_{60}$ Interestingly, the most probable $F_{\max}$ is about the same in both cases,

as well as Q_{60}, 12 ε/nm, but the shapes of the distributions differ: The distribution for $CATH_{60}$ is much broader than those of Q_{60} and V_{60}, evidencing the stronger compositional homogeneity in the latter two sets. The rougher look of the distribution for Q_{60} compared to V_{60} is likely due to the one order of magnitude smaller statistics. Furthermore, it should be noted that $CATH_{60}$ leads to the biggest number of situations with no force peaks, (49 ± 4) %; and V_{60} to the smallest, (19.4 ± 0.7) %. Despite the similarity of the distribution of the forces between Q_{60} and V_{60}, the geometrical character of structures in the two sets are distinct: V_{60} conformers are more compact and less elongated than Q_{60} or $CATH_{60}$. Furthermore, this figure indicates that size and shape of a chain need not be correlated.

Fig. 7.8 presents a scatter plot of the whole CATH database with the $CATH_{60}$ set highlighted in red. This plot shows how $\langle z \rangle$ varies from 5.5 to 8.5, and while large values of $F_{\max}$ arise for $\langle z \rangle$ between 6.3 and 8.1, many large $\langle z \rangle$ structures come with average or even small forces, including $F_{\max}$=0. Similarly, it also shows that large $\mathbb{SS}$ may come with low or zero forces and large $F_{\max}$ may arise when $\mathbb{SS}$ is at its lower range. This observation further proves that there is no correlation between $F_{\max}$ and $\langle z \rangle$ or $\mathbb{SS}$ not only in Q_n but also for general (folded) proteins. Moreover, the lower panels of Fig. 7.8 also prove that $F_{\max}$ is unrelated to hydrogen-bonded turns, α-helical and β-strand content. This last observation might be striking in the field, since it is typical for mostly-β proteins to present high forces. Nonetheless, situations may arise in which the cooperativity between the hydrogen bonds is not much and even an all-β protein might yield low $F_{\max}$.

7.4. Structural and dynamical analysis of β-amyloid

After the analysis of Q_n, we decided to apply the same type of analysis to Aβ and its mutants. This study included not only the three experimentally studied species $A\beta_{42}$, E22G $A\beta_{42}$ (arctic mutation) and F19S/L34P $A\beta_{42}$, but also three others: E22G/I31E $A\beta_{42}$, a mutant on the arctic mutation that reverts the gain in toxicity by accelerating the fibrillogenic pathway; the shorter wild type $A\beta_{40}$, which presents less toxicity than the longer peptide; and E3R $A\beta_{40}$, a mutant of the shorter peptide that induces a toxic gain-of-function [32].

Fig. 7.9 shows that all the Aβ conformations have similar size and shape, and they do not differ from Q_{40}. Fig. 7.10 presents the $F_{\max}$ histograms for

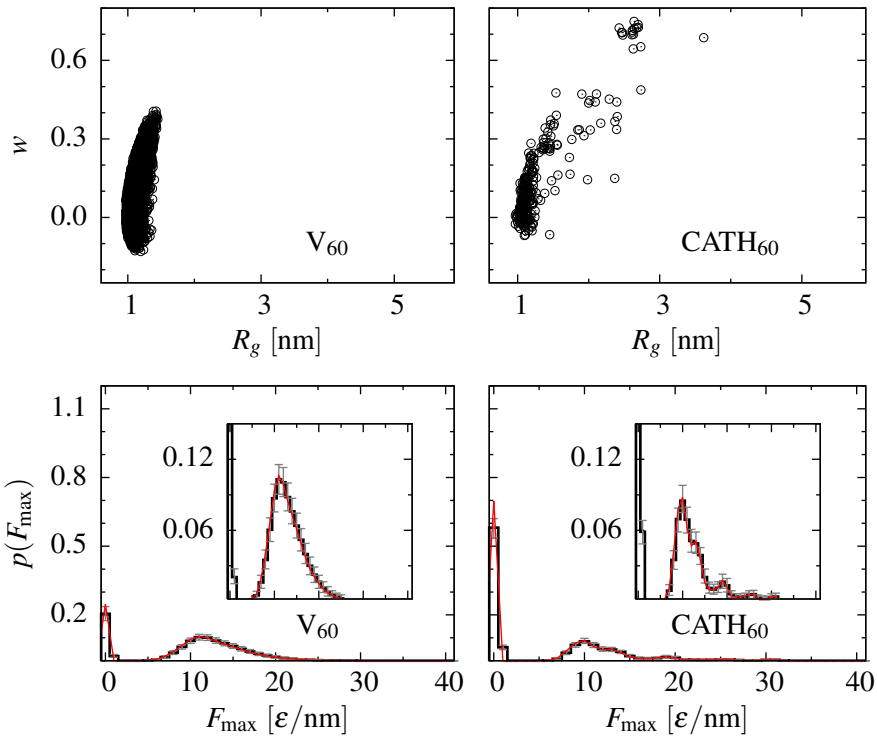


Figure 7.7: Analysis of V_{60} and $CATH_{60}$. The red curve in the F_{max} histograms represent the data as directly obtained, while the black bars mark a 95 % confidence interval.

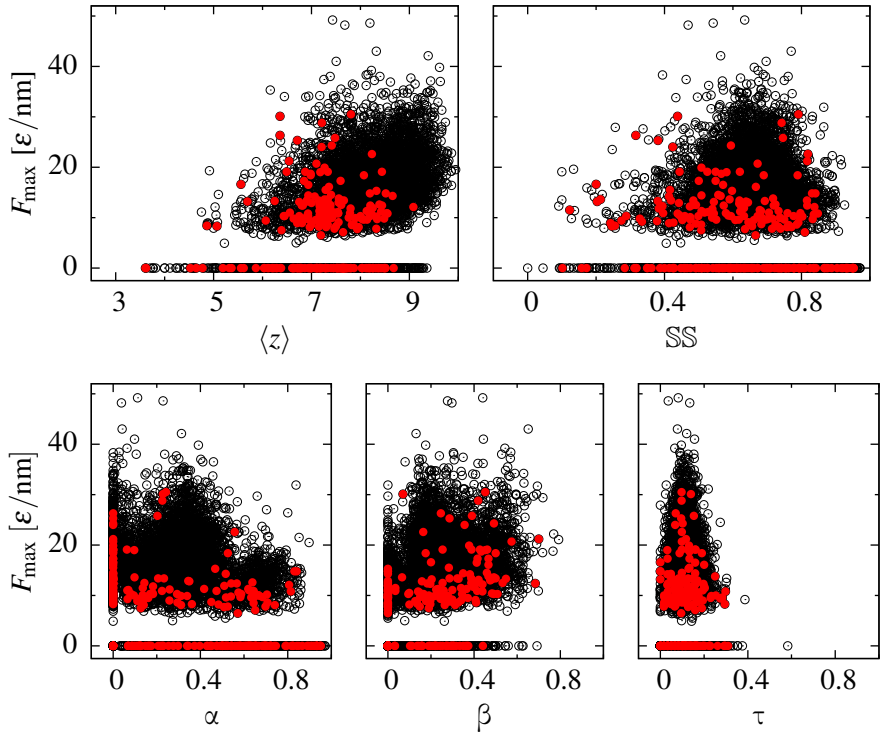


Figure 7.8: Dependence of F_{max} with structural descriptors. The descriptors are $\langle z \rangle$, $\mathbb{SS}$, α-helical content, β-strand content and hydrogen-bonded-turns content (τ). The black dots represent all proteins in CATH. The proteins with 57 to 63 residues ($CATH_{60}$) are highlighted in red. The scatter plots show that none of the descriptors, including β-strand content, are a good predictors for the F_{max} of a protein.

	$A\beta_{40}$	E3R	$A\beta_{42}$	Arc	I31E	DM
t [μs]	0.65	0.65	0.64	0.67	1.03	0.67
N	22	242	207	267	386	256
$F_{\max} > 0$ [%]	19±2	48±4	46±3	45±3	48±4	45±3
$\max(F_{\max})$ [ε/nm]	12	26	19	23	22	19
$\max(F_{\max})$ [pN]	132	286	209	253	242	209
$\langle z \rangle < \min(\langle z \rangle)$ [%]	36.8±0.6	15±3	15±3	22±3	15±3	18±3

Table 7.3: Dynamical parameters of Aβ. Table shows the simulation time, the number of independent conformers, the percentage of mechanostable ones, the maximum $F_{\max}$ and the number of volatile conformers (assessed by $\langle z \rangle$ being smaller than the stiff limit) for each studied Aβ. Arc stands for arctic mutation (E22G Aβ) and DM for Double Mutant (F19S/L34P Aβ). I31E is a mutation on Arc.

Aβ, which are similar to those of Q_n in shape. In Tab. 7.3 is a summary of the relevant parameters of these systems. It is interesting that the E3R mutation on $A\beta_{40}$ changes the landscape significantly, while the mutations on $A\beta_{42}$ peptide affect the $F_{\max}$ distribution very little, with close to no significant changes.

Given that similar statistics to Q_n in number of conformers were obtained in less simulation time, we conclude that the conformational polymorphism in Aβ is much richer. This observation correlates with the fact that, in the experiments, many more mechanostable conformers were found for Aβ than for Q_n. Knotted structures were not found in Aβ, which may be again due to the small sampling of the conformational space.

The temporal evolution of the structures was also studied for Aβ. Interestingly, the average lifetime of the mutants presented significant differences with those of the wild types as expected: As observed for Q_n, toxicity-inducing mutants tend to last longer than those that are less toxic. Fig. 7.11 shows the escape probability as a function of time, which is lower for the toxic mutants (E3R compared to $A\beta_{40}$ and Arc compared to $A\beta_{42}$) and is recovered in the non-toxic mutation I31E Arc. Interestingly, the non-fibrillogenic mutation F19S/L34P $A\beta_{42}$ shows no significant differences with the wild type, which leads to the reasoning that the former forms less stable conformers experimentally and does not aggregate due to the hidrophobicity change in the mutation, which does not affect the lifetime of the formed structures.

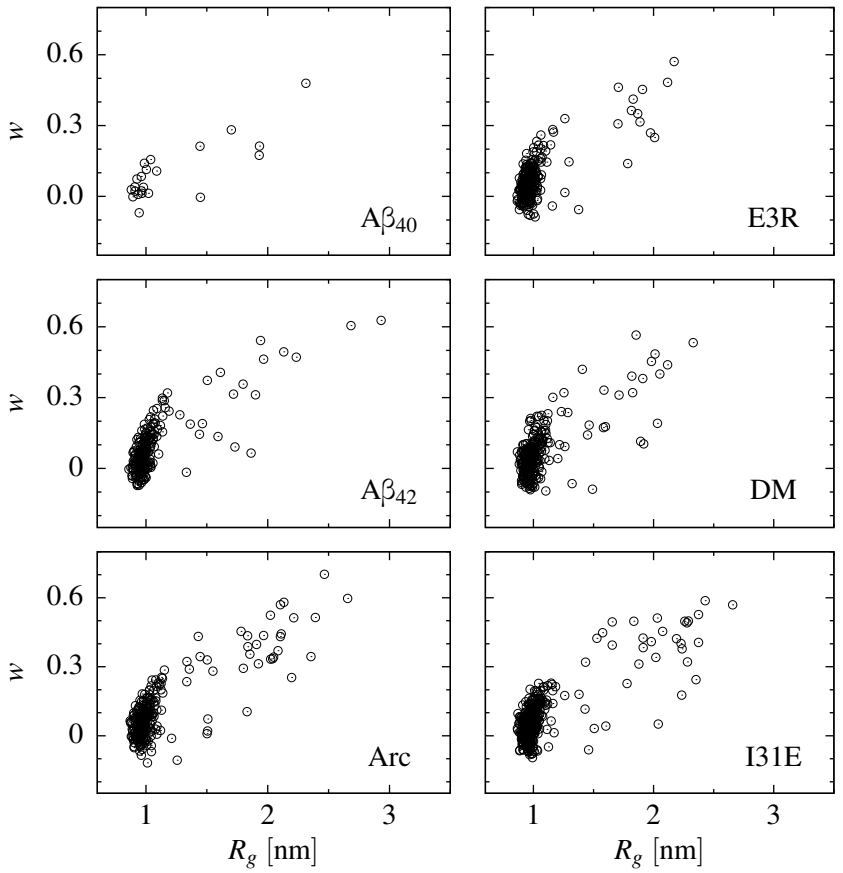


Figure 7.9: Structural characterization of Aβ. The graphs show the dispersion plot of R_g *vs.* w for the studied Aβ species. Arc stands for arctic mutation (E22G Aβ) and DM for Double Mutant (F19S/L34P Aβ). I31E is a mutation on Arc.

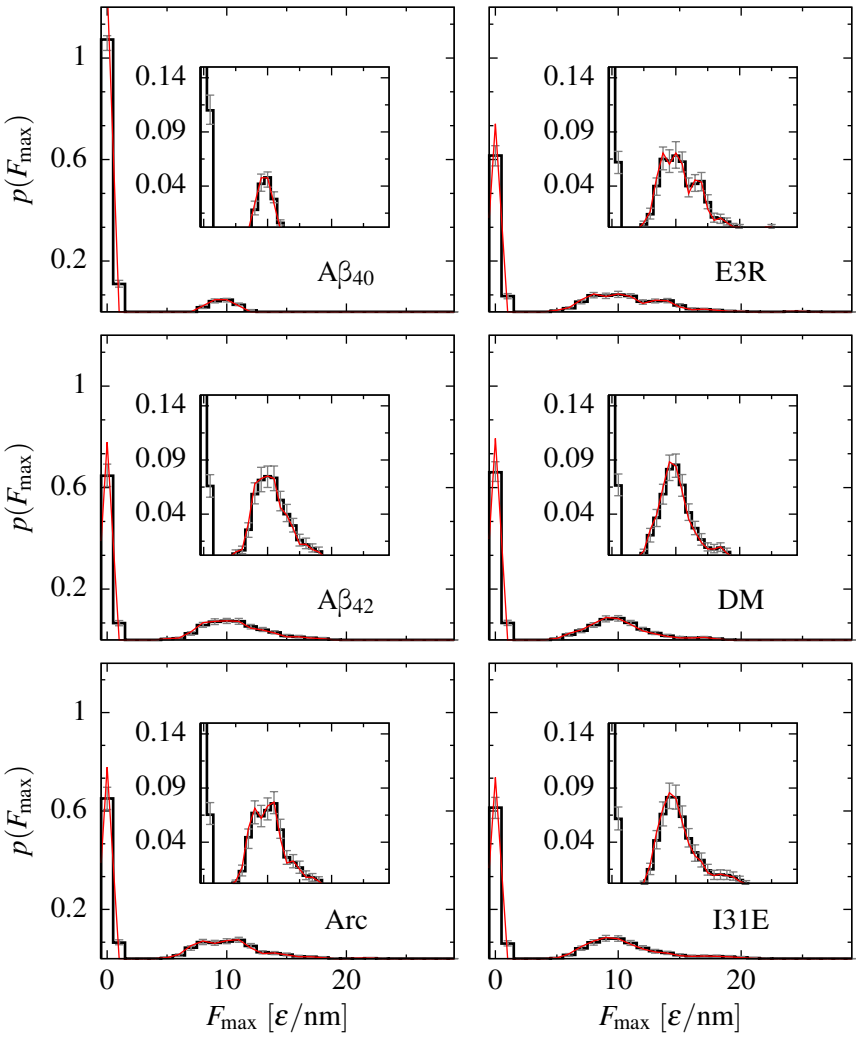


Figure 7.10: Distributions of F_{max} for Aβ. The red lines show the obtained distribution, while the black lines show a 95 % confidence interval. The insets show a close-up look at the distribution of forces for the mechanostable conformers. Arc stands for arctic mutation (E22G Aβ) and DM for Double Mutant (F19S/L34P Aβ). I31E is a mutation on Arc.

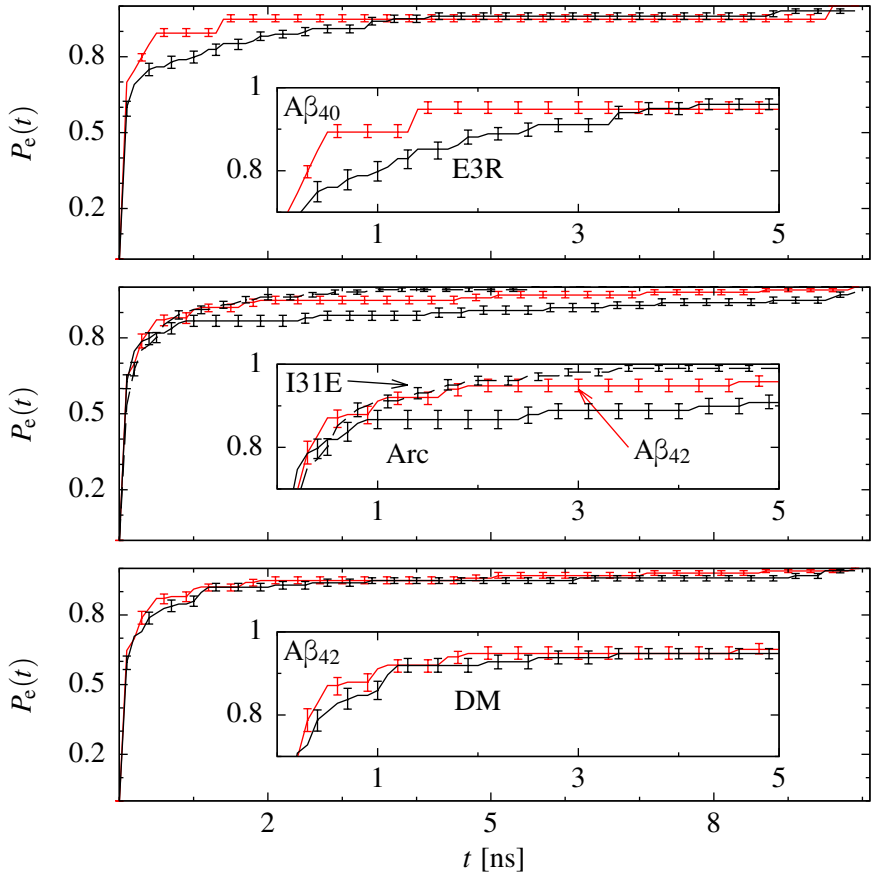


Figure 7.11: Time evolution of Aβ. For each of the studied species, 100 randomly chosen structures have been placed under a free-dynamics evolution for 10 ns. After that, the RMSD has been studied and the last time when it fluctuates above .2 nm is recorded as the residence time. The graph shows the escape probability ($P_e(t)$), defined as the probability of the residence time being smaller than t. The wild type species are plotted in red. The top graph compares $A\beta_{40}$ to its E3R mutation. The middle graph shows how the E22G (Arc) mutation on $A\beta_{42}$ induces longer lifetimes, and how I31E mutation on Arc reverts this change. The bottom panel compares $A\beta_{42}$ to its non-fibrillogenic mutant F19S/L34P (DM), with no significant differences.

7.5. Summary

We have explored a part of the energy landscape of Q_n for several chain lengths and Aβ with some critical mutations. We have studied their geometrical, structural and dynamical characteristics with the goal of finding a relationship between the conformations and the toxicity.

The discovery of knotted conformations of length 36, even if only in Q_{60}, leads to the interesting possibility that knotted conformations might be involved in toxicity by a proteasome blockade, supporting the conformational change hypothesis [31]. However, a greater exploration of the landscape of these neurotoxic proteins would be necessary to determine if these species are relevant for toxicity as the main toxic pathway or among many others.

The study of the temporal evolution presented another insight in the toxicity mechanisms, by which the toxic conformers would fluctuate more slowly than those that are innocuous. This result on monomeric species goes along the lines of recent experimental studies on the lifetime of oligomers, which indicate that toxic species last longer than non-toxic ones [151, 152].

Finally, the fact that the $F_{\max}$ behaves like an avalanche close to the disease threshold holds in itself yet another possibility for the toxicity being related to the resistance to unfolding of each of the conformers. Nonetheless, the values of $F_{\max}$ need to be taken carefully since these proteins fluctuate out of their conformation in short timescales and the *in vivo* pulling geometry is completely different than N-C direction typically used in an AFM or in simulations [101].

8. Proteasomal degradation of neurotoxic proteins

After finding that knots could hinder the degradation process and considering our mechanical hypothesis on neurodegeneration [31], we decided to explore the protein degradation machinery in this context. To that end, we studied the unfolding times of each of the conformers at several pulling forces using a recently developed model for a proteasome [101].

8.1. The model of the proteasome

The unfolding of the protein prior to degradation is done at the entrance of the proteasome, where a motor pulls the protein against a small pore. The model we used consists of a torus resting on a cylinder that together generate a pore-like structure, representing the entrance to the proteasome. There are two dimensions to know when building such a model: the size of the pore and the length of the cylinder. The latter lacks of importance in this case, since the interest relies only in the unfolding part, before the protein is transferred to the degradation chamber. The diameter of the channel, on the other hand, needs to be determined.

It was computed averaging the distances between the heavy atoms from the experimentally solved structure of a bacterial proteasome-like model, ClpX. The inner radius obtained was ≈ 0.7 nm [153, 154], a small-enough size to prevent the formation of secondary structure, but still allowing some movement of the chain in the cylinder.

The entrance of the proteasome is typically funnel-like, where polyubiquitin binding takes place in the degradation process. In this case, this structure has been modelled by choosing a torus-like structure and adjusting its radii to fit the experimental shape: 1.3 nm as the major radius and 0.6 nm as the minor. This sizes accommodate the protein at the entrance of the proteasome and generate enough surface to account for the binding places (which are not modelled), but presents an entrance of 0.7 nm at its narrowest point, ensuring the protein unfolding. In order to account for the space present in a proteasome between the narrow entrance pore and the comparatively wider degradation chamber, the radius of the cylinder after the torus was taken to be 0.8 nm.

The interaction between the protein and the surface of the torus as well as the inner surface of the cylindrical pore is assumed to be repulsive and given by the truncated Lennard-Jones potential as given by Eq. 8.1, where d_i is the closest distance between the ith AA and the surface of the torus. The distance $r_{\min} = 6$ nm is the coordinate of the minimum of the potential, which takes into account excluded volumes of residues and wall atoms. We take σ to be $0.5^{1/6} r_{\min} = 5.345$ nm.

$$V(d_i) = \begin{cases} 4\varepsilon \left[\left(\frac{\sigma}{d_i}\right)^{12} - \left(\frac{\sigma}{d_i}\right)^{6} \right] & , \quad d_i \leq r_{\min} \\ 0 & , \quad d_i > r_{\min} \end{cases} \tag{8.1}$$

The unfolding of a protein is generated by conformational changes on the entrance proteins that push the protein inside the pore. In this model, rotation and conformational changes cannot occur since the shape of the entrance is flat, so the traction of the protein into the pore is modelled by a force acting on the residue from which the protein is sucked. Although other cases can be envisioned, we assume C-terminal pulling in the simulations of this work, which in the case of knot untying would be the worst-case scenario in huntingtin due to the position of the Q_n in it – at the N-terminus. The pulling force value should be adjusted to match the degradation speed in experiments, and so it is expected to be small. Nonetheless, small forces yield large unfolding times that would be unreachable

in the computations. Thus, this model is used to study the dependence of the unfolding time with the pulling force.

It is important to note that the pulling from the proteasome occurs from one end while the other is free. This pulling geometry is significantly different from one where the proteasome pulls on one side while the other end is being restricted (*e.g.* by an external measuring device, even if it is not exerting a force). This is so because the constraint induces a shape restriction in which the protein is oriented and not free enough to move or probably even rotate, inducing changes in the mechanical clamps exposed to the unfolding region and, thus, changing the timescales for the protein unfolding. Therefore, experimental works such as Ref. [45] and [46] should be looked at with care.

The initial orientation of the protein might be an issue in the determination of the unfolding times. However, given that the entering is modelled with a force on one single residue, this model assumes the least invasive technique, where the pulled atom starts directly above and closest to the proteasome. The orientation of the rest of the protein is such that its N-C axis coincides with the symmetry axis of the proteasome.

8.2. Polyglutamine in the proteasome

Using the aforementioned model, we simulated the unfolding of the conformers obtained in Sec. 7 through the proteasome. We compared the pulling at a constant velocity with the one done AFM-like. Nonetheless, the proteasome is assumed to work with a periodic force pattern: An ATP molecule arrives diffusively to each subunit, and it is not until all of them have ATP bound that the hydrolysis is produced and the subunits perform the conformational change that exerts the force. Thus, we studied the more similar situation of constant force pulling.

8.2.1. Differences between AFM and proteasome pulling

After pulling each of the conformers along the N-C direction in what we call an AFM-like fashion, we performed constant speed pulling from the C-terminus against the pore while keeping the N-terminus free. This, as expected, changed the unfolding curves due to the difference in geometry. We decided on comparing Q_{20} to Q_{60} in this simulation, since they are well below and well above the disease threshold and must therefore show more clearly the differences between the conformations, if any. The results of this study are shown in Fig. 8.1.

Indeed, not only do the shape of the curves change, but also the shape of the $F_{\max}$ histograms. In particular, Q_{20} conformers seem to unfold much easier in the proteasome than in the AFM, shifting the $F_{\max}$ distribution to the left. Furthermore, Q_{60} structures present more resistance to unfolding through the proteasome than in the AFM, since even if the right shifting of the distribution is not statistically significant, the number of mechanostable conformers rose from (61 ± 3) % to (91 ± 3) %.

It is worth noticing that the mechanically resistant elements might be different in the proteasome than they are in the AFM. The right panels in Fig. 8.1 show that $F_{\max}$ through the proteasome and through the AFM are uncorrelated and conformers that unfold with no force peaks in one may have very high mechanical stability in the other, both in the cases of Q_{20} and Q_{60}.

8.2.2. Unfolding time of the conformers

After studying constant velocity pulling, we also performed the unfolding at constant force for different forces, ranging from 5 to 30 ε/nm (55 to 330 pN). For this study, we define the unfolding time ($t_{\max}$) as the time it takes for the protein to be completely inside of the proteasome chamber. This simulations take longer than constant velocity ones, and are limited by the software to a maximum of 10^7 $\tau \approx 10$ ms. The proteins that are not unfolded after this time can still be unfolded at longer timescales if the forces are low, but they are considered to be stalling the proteasome for higher forces. The stalling of the proteasome is normally due to some steric constraint such as a knot.

Fig. 8.2 compares the times in which knotted and unknotted proteins are unfolded in the proteasome, and the probability of proteasome stalling in each case. The findings show how the probability of stalling is not significant for high forces in the case of unknotted proteins, while it is between 10 % and 30 % for the knotted ones. Furthermore, in the case of small forces, the stalling probability grows both for the knotted and the unknotted, which might be attributed to certain resistance elements that take longer to unfold than the allowed simulation time, but the knotted conformations still stall with a significantly higher probability, sometimes of 80 %.

Even in the cases where the knotted structures have low stalling probability, the unfolding time of the knotted proteins is always significantly higher than that of their unknotted partners at the same force. This is especially so at low forces, where knotted conformers are not unfolded for forces below 10 ε/nm

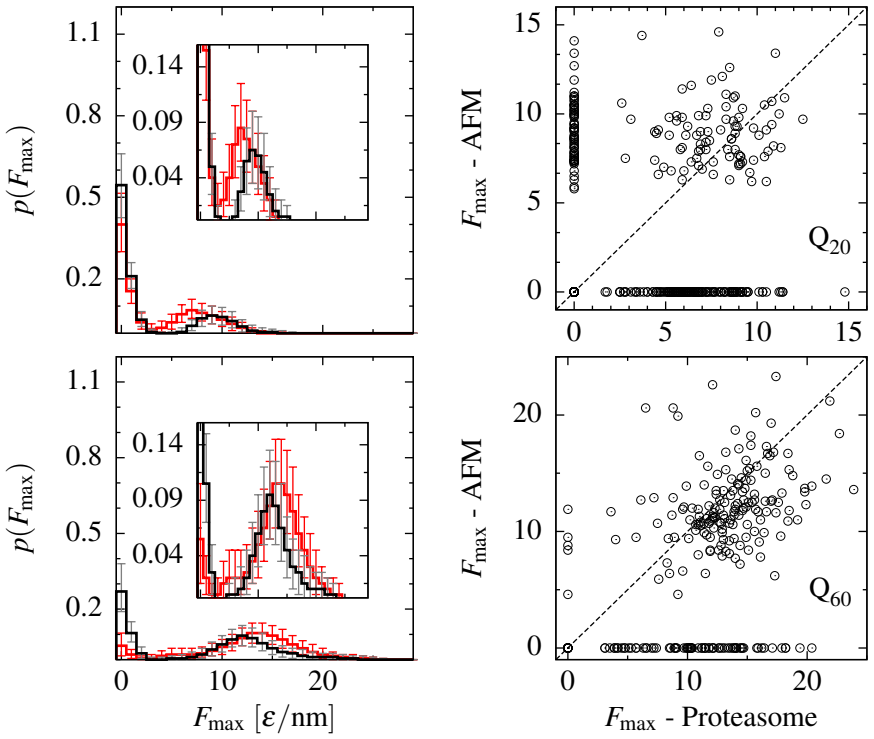


Figure 8.1: Comparison between AFM and proteasome pulling. The histograms on the left compare the proteasomal pulling (red) to the AFM-like stretching (black). The scatter plot on the right shows that the forces are not correlated: highly resistant conformers in the AFM may unfold easily in the proteasome and *vice versa*.

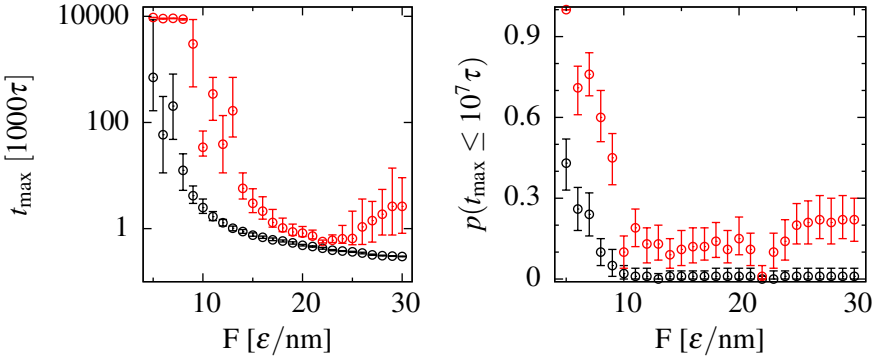


Figure 8.2: Effect of the knots in the proteasome. The left panel shows the median unfolding time for the unknotted (black) and knotted (red) conformers as a function of the pulling force. The right panel displays the probability of the proteasome stalling *vs.* the pulling force, again for the unknotted (black) and knotted (red) conformers. The error bars are a 95 % confidence interval.

while unknotted ones can be unfolded down to 5 ε/nm; and also for high forces, at which the knot is tightened at a high speed and is not free enough to easily rearrange and slip backwards toward the end of the chain. In the central range of forces, the knot is commonly untied, but degradation is still slower, which may act as a hindrance of the function of the degradation machinery even without the stalling.

8.3. Summary

We have used a proteasome model to study the differences between this more realistic case and the AFM pulling, using Q_n as the protein to unfold. The result is that the disease-related Q_{60} is even harder to degrade in the proteasome than in the AFM, while the non-disease-related Q_{20} is easier when pulled at constant speed.

We studied the degradation time when pulling at a constant force. We proved that knotted structures can sometimes stall the proteasome and, even if they do not, they present unfolding timescales that are longer than unknotted conformers at the same force. This finding suggests that even if a knotted conformation does not necessarily stall the proteasome, it may hinder its working for a longer time

than necessary and thus have a toxicity effect on the cell just from the delay in its degradation.

Part IV

Conclusions

Conclusions

1. We have explored the use of the Host–Guest strategy for Single Molecule Force Spectroscopy as a mechanical protection method in the whole range of mechanical stability and found that preferably the Host should unfold in a single step and with higher mechanical stability than the single-molecule markers.

2. We have studied the advantage of using chemical – as well structural – information in the computation of the contact map of a protein in the context of folding and unfolding in a structure-based model of molecular dynamics, and concluded that the combination of the structural model based on overlaps with the few number of contacts found by the Repulsion CSU algorithm is the safest option, even if it is in general comparable to simply using the overlap method.

3. We have experimenally analyzed by AFM the conformational space of polyglutamine expansions with different lengths and β-amyloid in its wild type form and some mutants, and discovered a rich mechanical polymorphism in terms of mechanical stability of the molecules that relates positively with the toxicity of the studied species and is inhibited – for polyglutamines – with the incubation with the antiamyloidogenic peptide QBP1.

4. We have delved further into the conformational space of polyglutamine and β-amyloid *in silico* by generating several possible conformations and studying their geometrical, structural and dynamic characteristics, which correlate well with the experimental results.

5. We propose three toxicity-generating mechanisms based on the dynamic studies on polyglutamine and β-amyloid: The slower fluctuations of the toxic species compared to the less- or non-toxic ones might induce failure in their biological function; the presence of conformers with high mechanical stability may hinder the degradation machinery; and the presence of knotted conformations may stop the degradation by stalling the cellular unfolding machinery.

6. We have probed *in silico* the proteasomal degradation of polyglutamine tracts and observed that the mechanical stability is not preserved or related

to the one obtained in an AFM-like scenario, as suggested by the different pulling geometry.

7. Finally, we have found that the knotted polyglutamine conformations sometimes induce a stalling of the proteasome, especially at high forces, and even when they do not, they increase the degradation time with respect to unknotted conformers at the same force, supporting the hypothesis of toxicity induced by blocking the unfolding machinery of the cell.

Conclusiones

1. Hemos explorado el uso de la estrategia Huésped–Invitado para Espectroscopía de Fuerza Monomolecular en todo el rango de estabilidades mecánicas y hemos encontrado que la protección mecánica es un requerimiento, y preferentemente el Huésped debe desplegarse a través de un solo evento y tener estabilidad mecánica mayor que la de los marcadores de monomolecularidad.

2. Hemos estudiado las ventajas del uso de información de origen químico – además de estructural – en el cálculo del mapa de contactos de una proteína en el contexto de plegamiento y desplegamiento bajo el modelo de dinámica molecular basado en estructura, y hemos concluido que la combinación del modelo estructural basado en superposiciones y en los pocos contactos encontrados por el algoritmo CSU de Repulsión es la opción más segura, si bien es en general comparable a usar únicamente el método de superposiciones.

3. Hemos analizado experimentalmente mediante AFM el espacio conformacional de las expansiones de poliglutamina de distintas longitudes y del β-amiloide en su forma silvestre y algunas de sus mutaciones, y descubierto un rico polimorfismo en términos de estabilidad mecánica de las moléculas que está positivamente relacionado con la toxicidad de las especies estudiadas y se inhibe – en el caso de las poliglutaminas – tras la incubación con el péptido antiamiloidogénico QBP1.

4. Hemos profundizado en el espacio conformacional de las poliglutaminas y de β-amiloide *in silico* generando multitud de conformaciones posibles y estudiando sus características geométricas, estructurales y dinámicas, que correlacionan con los resultados experimentales.

5. Proponemos tres mecanismos generadores de toxicidad basados en los estudios dinámicos de las poliglutaminas y β-amiloide: Las fluctuaciones de las especies tóxicas, más lentas que las de las inocuas, podrían inducir fallos en sus funciones biológicas; la presencia de moléculas con alta estabilidad mecánica podría enlentecer la maquinaria de degradación; y la presencia de conformaciones con nudos podría obstruir la degradación bloqueando la maquinaria de desplegado de la célula.

6. Hemos examinado *in silico* la degradación a través del proteasoma de tractos de poliglutamina y hemos observado que la estabilidad mecánica no se conserva ni está relacionada con la obtenida en un escenario tipo AFM, tal y como sugiere la distinta geometría de estiramiento.

7. Finalmente, hemos encontrado que las conformaciones de poliglutamina que presentan nudos inducen en ocasiones al bloqueo del proteasoma, especialmente a fuerzas altas, y aún cuando no lo hacen, provocan un aumento del tiempo de degradación con respecto a confórmeros no anudados a la misma fuerza, apoyando la hipótesis de toxicidad inducida por el bloqueo de la maquinaria celular de desplegado.

Bibliography

[1] Watson JD & Crick FH. *Molecular structure of nucleic acids*. Nature 1953. 171(4356):737–738.

[2] Pauling L, Corey RB & Branson HR. *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain*. Proceedings of the National Academy of Sciences USA 1951. 37(4):205–211.

[3] Sela M, White Jr FH & Anfinsen CB. *Reductive cleavage of disulfide bridges in ribonuclease*. Science 1957. 125(3250):691–692.

[4] Anfinsen CB *et al. Principles that govern the folding of protein chains*. Science 1973. 181(4096):223–230.

[5] Levinthal C. *Are there pathways for protein folding?* Journal de Chimie Physique 1968. 65(1):44–45.

[6] Krishna MM, Lin Y & Englander SW. *Protein misfolding: Optional barriers, misfolded intermediates, and pathway heterogeneity*. Journal of Molecular Biology 2004. 343(4):1095–1109.

[7] Fersht AR. *Nucleation mechanisms in protein folding*. Current Opinion in Structural Biology 1997. 7(1):3–9.

[8] Ptitsyn O & Rashin A. *A model of myoglobin self-organization*. Biophysical Chemistry 1975. 3(1):1–20.

[9] Karplus M & Weaver DL. *Protein-folding dynamics*. Nature 1976. 260:404–406.

[10] Dill KA. *Theory for the folding and stability of globular proteins*. Biochemistry 1985. 24(6):1501–1509.

[11] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM & Obradovic Z. *Intrinsic disorder and protein function*. Biochemistry 2002. 41(21):6573–6582.

[12] Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, Dyson HJ, Forman-Kay J, Fuxreiter M, Gsponer J, Han KH, Jones DT, Longhi S, Metallo SJ, Nishikawa K, Nussinov R, Obradovic Z, Pappu RV, Rost B, Selenko P, Subramaniam V, Sussman JL, Tompa P & Uversky VN. *What's in a name? Why these proteins are intrinsically disordered*. Intrinsically Disordered Proteins 2013. 1(1):e24 157. doi:10.4161/idp.24157.

[13] Bussell R & Eliezer D. *Residual structure and dynamics in Parkinson's disease-associated mutants of α-synuclein*. Journal of Biological Chemistry 2001. 276(49):45 996–46 003.

[14] Dunker AK & Obradovic Z. *The protein trinity-linking function and disorder*. Nature Biotechnology 2001. 19(9):805–806.

[15] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF & Jones DT. *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life*. Journal of Molecular Biology 2004. 337(3):635–645.

[16] Kopito RR & Ron D. *Conformational disease*. Nature Cell Biology 2000. 2(11):E207–E209.

[17] Uversky VN, Oldfield CJ & Dunker AK. *Intrinsically disordered proteins in human diseases: introducing the D2 concept*. Annual Review of Biophysics 2008. 37:215–246.

[18] Magrane M & Consortium U. *UniProt knowledgebase: a hub of integrated protein data*. Database 2011. 2011. doi:10.1093/database/bar009.

[19] Nasir J, Floresco SB, O'Kusky JR, Diewert VM, Richman JM, Zeisler J, Borowski A, Marth JD, Phillips AG & Hayden MR. *Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes*. Cell 1995. 81(5):811 – 823. ISSN 0092-8674. doi:http://dx.doi.org/10.1016/0092-8674(95)90542-1.

[20] Zuccato C, Ciammola A, Rigamonti D, Leavitt BR, Goffredo D, Conti L, MacDonald ME, Friedlander RM, Silani V, Hayden MR, Timmusk T, Sipione S & Cattaneo E. *Loss of huntingtin-mediated BDNF gene transcription in Huntington's disease*. Science 2001. 293(5529):493–498. doi:10.1126/science.1059581.

[21] Velier J, Kim M, Schwarz C, Kim TW, Sapp E, Chase K, Aronin N & DiFiglia M. *Wild-type and mutant huntingtins function in vesicle trafficking in the secretory and endocytic pathways*. Experimental Neurology 1998. 152(1):34 – 40. ISSN 0014-4886. doi:http://dx.doi.org/10.1006/exnr.1998.6832.

[22] *World Health Organization (WHO)*. `www.who.int`. Accessed: 2015-05-30.

[23] Petruska J, Hartenstine MJ & Goodman MF. *Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease*. Journal of Biological Chemistry 1998. 273(9):5204–5210. doi:10.1074/jbc.273.9.5204.

[24] Nagai Y, Fujikake N, Ohno K, Higashiyama H, Popiel HA, Rahadian J, Yamaguchi M, Strittmatter WJ, Burke JR & Toda T. *Prevention of polyglutamine oligomerization and neurodegeneration by the peptide inhibitor QBP1 in Drosophila*. Human Molecular Genetics 2003. 12(11):1253–1259. doi:10.1093/hmg/ddg144.

[25] Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJ, McFarlane HT *et al. Atomic structures of amyloid cross-β spines reveal varied steric zippers*. Nature 2007. 447(7143):453–457.

[26] Kayed R, Head E, Thompson JL, McIntire TM, Milton SC, Cotman CW & Glabe CG. *Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis*. Science 2003. 300(5618):486–489. doi:10.1126/science.1079469.

[27] Rajendran L & Annaert W. *Membrane trafficking pathways in Alzheimer's disease*. Traffic 2012. 13(6):759–770.

[28] Blennow K, de Leon MJ & Zetterberg H. *Alzheimer's disease*. The Lancet XXXX. 368(9533):387–403. doi:10.1016/S0140-6736(06)69113-7.

[29] Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, Hall K, Hasegawa K, Hendrie H, Huang Y, Jorm A, Mathers C, Menezes PR, Rimmer E & Scazufca M. *Global prevalence of dementia: a Delphi consensus study*. The Lancet XXXX. 366(9503):2112–2117. doi: 10.1016/S0140-6736(05)67889-0.

[30] Goldgaber D, Lerman M, McBride W, Saffiotti U & Gajdusek D. *Isolation, characterization, and chromosomal localization of human brain cDNA clones coding for the precursor of the amyloid of brain in Alzheimer's disease, Down's syndrome and aging.* Journal of Neural Transmission Supplementum 1986. 24:23–28.

[31] Hervás R, Oroz J, Galera-Prat A, Goñi O, Valbuena A, Vera AM, Gómez-Sicilia À, Losada-Urzáiz F, Uversky VN, Menéndez M *et al. Common features at the start of the neurodegeneration cascade*. PLoS Biology 2012. 10(5):e1001 335.

[32] Brorsson AC, Bolognesi B, Tartaglia GG, Shammas SL, Favrin G, Watson I, Lomas DA, Chiti F, Vendruscolo M, Dobson CM *et al. Intrinsic determinants of neurotoxic aggregate formation by the amyloid β peptide*. Biophysical Journal 2010. 98(8):1677–1684.

[33] Tycko R. *Solid state NMR studies of amyloid fibril structure*. Annual Review of Physical Chemistry 2011. 62:279.

[34] Hoel P. *Introduction to mathematical statistics*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley, 1984. ISBN 9780471890454.

[35] Ferreon ACM, Moran CR, Gambin Y & Deniz AA. *Single-molecule fluorescence studies of intrinsically disordered proteins*. Methods in Enzymology 2010. 472:179–204.

[36] Linke WA, Ivemeyer M, Olivieri N, Kolmerer B, Rüegg CJ & Labeit S. *Towards a molecular understanding of the elasticity of titin*. Journal of Molecular Biology 1996. 261(1):62–71.

[37] Vale RD. *The molecular motor toolbox for intracellular transport*. Cell 2003. 112(4):467 – 480. ISSN 0092-8674. doi:http://dx.doi.org/10.1016/S0092-8674(03)00111-9.

[38] Muramatsu Y, Kamegai A, Shiba T, Shrestha P, Takai Y, Mori M, Ilg E, Schafer B & Heizmann C. *Histochemical characteristics of calcium binding S100 proteins and bone morphogenetic proteins in chondro-osseous tumors*. Oncology Reports 1997. 4(1):49–53.

[39] Lehman IR, Bessman MJ, Simms ES & Kornberg A. *Enzymatic synthesis of deoxyribonucleic acid: I. Preparation of substrates and partial purification of an enzyme from Escherichia Coli*. Journal of Biological Chemistry 1958. 233(1):163–170.

[40] Zhang S, Udho E, Wu Z, Collier RJ & Finkelstein A. *Protein translocation through anthrax toxin channels formed in planar lipid bilayers*. Biophysical Journal 2004. 87(6):3842–3849.

[41] Lecker SH, Goldberg AL & Mitch WE. *Protein degradation by the ubiquitin-proteasome pathway in normal and disease states*. Journal of the American Society of Nephrology 2006. 17(7):1807–1819. doi:10.1681/ASN.2006010083.

[42] Sauer RT, Bolon DN, Burton BM, Burton RE, Flynn JM, Grant RA, Hersch GL, Joshi SA, Kenniston JA, Levchenko I *et al.* *Sculpting the proteome with AAA+ proteases and disassembly machines*. Cell 2004. 119(1):9–18.

[43] Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K & Walter P. *Molecular biology of the cell*. 500 Tips Series. Taylor & Francis, 2014. ISBN 9780815344643.

[44] Coux O, Tanaka K & Goldberg AL. *Structure and functions of the 20S and 26S proteasomes*. Annual Review of Biochemistry 1996. 65(1):801–847.

[45] Maillard RA, Chistol G, Sen M, Righini M, Tan J, Kaiser CM, Hodges C, Martin A & Bustamante C. *ClpX (P) generates mechanical force to unfold and translocate its protein substrates*. Cell 2011. 145(3):459–469.

[46] Aubin-Tam ME, Olivares AO, Sauer RT, Baker TA & Lang MJ. *Single-molecule protein unfolding and translocation by an ATP-fueled proteolytic machine*. Cell 2011. 145(2):257–267.

[47] Sauer RT & Baker TA. *AAA+ proteases: ATP-fueled machines of protein destruction*. Annual Review of Biochemistry 2011. 80:587–612.

[48] Baró AM, Miranda R, Alamán J, García N, Binnig G, Rohrer H, Gerber C & Carrascosa JL. *Determination of surface topography of biological specimens at high resolution by scanning tunnelling microscopy*. Nature 1985. 315(6016):253–254.

[49] Binnig G, Quate CF & Gerber C. *Atomic force microscope*. Physical Review Letters 1986. 56(9):930–933.

[50] Francis LW, Lewis PD, Wright CJ & Conlan RS. *Atomic force microscopy comes of age.* Biology of the Cell 2010. 102(2):133–143. doi:10.1042/BC20090127.

[51] Rief M, Gautel M, Oesterhelt F, Fernandez JM & Gaub HE. *Reversible unfolding of individual titin immunoglobulin domains by AFM.* Science 1997. 276(5315):1109–1112.

[52] Kellermayer MS, Smith SB, Granzier HL & Bustamante C. *Folding-unfolding transitions in single titin molecules characterized with laser tweezers*. Science 1997. 276(5315):1112–1116.

[53] Hutter JL & Bechhoefer J. *Calibration of atomicforce microscope tips*. Review of Scientific Instruments 1993. 64(7):1868–1873. doi:http://dx.doi.org/10.1063/1.1143970.

[54] Martin Y, Williams C & Wickramasinghe HK. *Atomic force microscope–force mapping and profiling on a sub 100-Å scale*. Journal of Applied Physics 1987. 61(10):4723–4729.

[55] García R, Magerle R & Perez R. *Nanoscale compositional mapping with gentle forces*. Nature Materials 2007. 6(6):405–411.

[56] De Pablo P, Colchero J, Gomez-Herrero J & Baro A. *Jumping mode scanning force microscopy*. Applied Physics Letters 1998. 73(22):3300–3302.

[57] Horcas I, Fernández R, Gomez-Rodriguez J, Colchero J, Gómez-Herrero J & Baro A. *WSXM: a software for scanning probe microscopy and a tool for nanotechnology*. Review of Scientific Instruments 2007. 78(1):013 705.

[58] Valbuena A, Oroz J, Vera AM, Gimeno A, Gómez-Herrero J & Carrión-Vázquez M. *Quasi-simultaneous imaging/pulling analysis of single polyprotein molecules by atomic force microscopy*. Review of Scientific Instruments 2007. 78(11):113 707.

[59] Bustamante C, Marko J, Siggia E & Smith S. *Entropic elasticity of lambda-phage DNA*. Science 1994. 265(5178):1599–1600. doi:10.1126/science.8079175.

[60] Li H, Oberhauser AF, Redick SD, Carrion-Vazquez M, Erickson HP & Fernandez JM. *Multiple conformations of PEVK proteins detected by single-molecule techniques*. Proceedings of the National Academy of Sciences USA 2001. 98(19):10 682–10 686.

[61] Oberhauser AF, Marszalek PE, Erickson HP & Fernandez JM. *The molecular elasticity of the extracellular matrix protein tenascin*. Nature 1998. 393(6681):181–185.

[62] Humphrey W, Dalke A & Schulten K. *VMD: visual molecular dynamics*. Journal of Molecular Graphics 1996. 14(1):33–38.

[63] Dietz H, Bertz M, Schlierf M, Berkemeier F, Bornschlogl T, Junker JP & Rief M. *Cysteine engineering of polyproteins for single-molecule force spectroscopy*. Nature Protocols 2006. 1(1):80–84. ISSN 1754-2189. doi:10.1038/nprot.2006.12.

[64] Marszalek PE, Lu H, Li H, Carrion-Vazquez M, Oberhauser AF, Schulten K & Fernandez JM. *Mechanical unfolding intermediates in titin modules*. Nature 1999. 402(6757):100–103.

[65] Wiita AP, Ainavarapu SRK, Huang HH & Fernandez JM. *Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques*. Proceedings of the National Academy of Sciences USA 2006. 103(19):7222–7227.

[66] Wiita AP, Perez-Jimenez R, Walther KA, Gräter F, Berne B, Holmgren A, Sanchez-Ruiz JM & Fernandez JM. *Probing the chemistry of thioredoxin catalysis with force*. Nature 2007. 450(7166):124–127.

[67] Grützner A, Garcia-Manyes S, Kötter S, Badilla CL, Fernandez JM & Linke WA. *Modulation of titin-based stiffness by disulfide bonding in the cardiac titin N2-B unique sequence*. Biophysical Journal 2009. 97(3):825–834.

[68] Sosnick T. *Comment on "Force-clamp spectroscopy monitors the folding trajectory of a single protein"*. Science 2004. 306(5695):411–411.

[69] Best RB & Hummer G. *Comment on "Force-clamp spectroscopy monitors the folding trajectory of a single protein"*. Science 2005. 308(5721):498. doi:10.1126/science.1106969.

[70] Garcia-Manyes S, Brujić J, Badilla CL & Fernández JM. *Force-clamp spectroscopy of single-protein monomers reveals the individual unfolding and folding pathways of I27 and ubiquitin*. Biophysical Journal 2007. 93(7):2436–2446.

[71] Li H, Linke WA, Oberhauser AF, Carrion-Vazquez M, Kerkvliet JG, Lu H, Marszalek PE & Fernandez JM. *Reverse engineering of the giant muscle protein titin*. Nature 2002. 418(6901):998–1002.

[72] Steward A, Toca-Herrera JL & Clarke J. *Versatile cloning system for construction of multimeric proteins for use in atomic force microscopy*. Protein Sciences 2002. 11(9):2179–2183.

[73] Peng Q & Li H. *Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: toward engineering multifunctional elastomeric proteins*. Journal of the American Chemical Society 2009. 131(39):14 050–14 056.

[74] Oroz J, Hervás R & Carrión-Vázquez M. *Unequivocal single-molecule force spectroscopy of proteins by AFM using pFS vectors*. Biophysical Journal 2012. 102(3):682–690.

[75] Finley D, Bartel B & Varshavsky A. *The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis.* Nature 1989. 338(6214):394–401.

[76] Shannon CE. *Communication in the presence of noise*. Proceedings of the IEEE 1984. 72(9):1192–1201.

[77] Ng SP, Randles LG & Clarke J. *Single molecule studies of protein folding using atomic force microscopy*. In *Protein Folding Protocols*, pages 139–167. Springer, 2006.

[78] Evans E & Ritchie K. *Dynamic strength of molecular adhesion bonds*. Biophysical Journal 1997. 72(4):1541–1555.

[79] Dudko OK, Hummer G & Szabo A. *Intrinsic rates and activation free energies from single-molecule pulling experiments*. Physical Review Letters 2006. 96(10):108 101.

[80] Rico F, Gonzalez L, Casuso I, Puig-Vidal M & Scheuring S. *High-speed force spectroscopy unfolds titin at the velocity of molecular dynamics simulations*. Science 2013. 342(6159):741–743.

[81] Schlierf M, Berkemeier F & Rief M. *Direct observation of active protein folding using lock-in force spectroscopy*. Biophysical Journal 2007. 93(11):3989–3998.

[82] Kessler M, Gottschalk KE, Janovjak H, Muller DJ & Gaub HE. *Bacteriorhodopsin folds into the membrane against an external force*. Journal of Molecular Biology 2006. 357(2):644–654.

[83] Kim M, Abdi K, Lee G, Rabbi M, Lee W, Yang M, Schofield CJ, Bennett V & Marszalek PE. *Fast and forceful refolding of stretched α-helical solenoid proteins*. Biophysical Journal 2010. 98(12):3086–3092.

[84] Oberhauser AF, Hansma PK, Carrion-Vazquez M & Fernandez JM. *Stepwise unfolding of titin under force-clamp atomic force microscopy*. Proceedings of the National Academy of Sciences USA 2001. 98(2):468–472.

[85] Fernandez JM & Li H. *Force-clamp spectroscopy monitors the folding trajectory of a single protein*. Science 2004. 303(5664):1674–1678.

[86] Aioanei D, Lv S, Tessari I, Rampioni A, Bubacco L, Li H, Samorì B & Brucale M. *Single-molecule-level evidence for the osmophobic effect*. Angewandte Chemie International Edition 2011. 50(19):4394–4397.

[87] Koti Ainavarapu SR, Wiita AP, Dougan L, Uggerud E & Fernandez JM. *Single-molecule force spectroscopy measurements of bond elongation during a bimolecular reaction*. Journal of the American Chemical Society 2008. 130(20):6479–6487.

[88] Brujić J, Walther KA, Fernandez JM *et al. Single-molecule force spectroscopy reveals signatures of glassy dynamics in the energy landscape of ubiquitin*. Nature Physics 2006. 2(4):282–286.

[89] Kuo TL, Garcia-Manyes S, Li J, Barel I, Lu H, Berne BJ, Urbakh M, Klafter J & Fernández JM. *Probing static disorder in Arrhenius kinetics by single-molecule force spectroscopy*. Proceedings of the National Academy of Sciences USA 2010. 107(25):11 336–11 340.

[90] Garcia-Manyes S, Dougan L, Badilla CL, Brujić J & Fernández JM. *Direct observation of an ensemble of stable collapsed states in the mechanical folding of ubiquitin*. Proceedings of the National Academy of Sciences USA 2009. 106(26):10 534–10 539.

[91] Berkovich R, Garcia-Manyes S, Klafter J, Urbakh M & Fernández JM. *Hopping around an entropic barrier created by force*. Biochemical and Biophysical Research Communications 2010. 403(1):133–137.

[92] Berkovich R, Garcia-Manyes S, Urbakh M, Klafter J & Fernandez JM. *Collapse dynamics of single proteins extended by force*. Biophysical Journal 2010. 98(11):2692–2701.

[93] Ritort F. *Single-molecule experiments in biological physics: methods and applications*. Journal of Physics: Condensed Matter 2006. 18(32):R531.

[94] Valbuena A, Oroz J, Vera AM, Gimeno A, Gómez-Herrero J & Carrión-Vázquez M. *Quasi-simultaneous imaging/pulling analysis of single polyprotein molecules by atomic force microscopy*. Review of Scientific Instruments 2007. 78(11):113 707.

[95] Alder BJ & Wainwright T. *Studies in molecular dynamics. I. General method*. The Journal of Chemical Physics 1959. 31(2):459–466.

[96] Allen P & Tildesley D. *Computer simulation of liquids*. Oxford Science Publ. Clarendon Press, 1989. ISBN 9780198556459.

[97] Chandler D. *Equilibrium theory of polyatomic fluids*. Studies in Statistical Mechanics 1982. 8:274.

[98] Goldstein H. *Classical mechanics*. Addison-Wesley series in physics. Addison-Wesley Publishing Company, 1980. ISBN 9780201029185.

[99] Shaw DE, Dror RO, Salmon JK, Grossman J, Mackenzie KM, Bank JA, Young C, Deneroff MM, Batson B, Bowers KJ *et al. Millisecond-scale molecular dynamics simulations on Anton*. In *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*. IEEE, 2009 pages 1–11.

[100] Lu H, Isralewitz B, Krammer A, Vogel V & Schulten K. *Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation*. Biophysical Journal 1998. 75(2):662–671.

[101] Wojciechowski M, Szymczak P, Carrión-Vázquez M & Cieplak M. *Protein unfolding by biological unfoldases: insights from modeling*. Biophysical Journal 2014. 107(7):1661–1668.

[102] Piana S & Laio A. *A bias-exchange approach to protein folding.* Journal of Physical Chemistry B 2007. 111(17):4553–4559. doi:10.1021/jp0678731.

[103] Sugita Y & Okamoto Y. *Replica-exchange molecular dynamics method for protein folding*. Chemical Physics Letters 1999. 314(1):141–151.

[104] Laio A & Parrinello M. *Escaping free-energy minima*. Proceedings of the National Academy of Sciences 2002. 99(20):12 562–12 566.

[105] Bussi G, Gervasio FL, Laio A & Parrinello M. *Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics*. Journal of the American Chemical Society 2006. 128(41):13 435–13 441. doi:10.1021/ja062463w. PMID: 17031956.

[106] Pietrucci F & Laio A. *A collective variable for the efficient exploration of protein beta-sheet structures: Application to SH3 and GB1*. Journal of Chemical Theory and Computation 2009. 5(9):2197–2201.

[107] Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R *et al.* *New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures*. Nucleic Acids Research 2012. page gks1211.

[108] Abe H & Gō N. *Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins*. Biopolymers 1981. 20(5):1013–1031.

[109] Veitshans T, Klimov D & Thirumalai D. *Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties*. Folding and Design 1997. 2(1):1–22.

[110] Sułkowska JI & Cieplak M. *Selection of optimal variants of Gō-like models of proteins through studies of stretching*. Biophysical Journal 2008. 95(7):3174–3191.

[111] Kwiecińska JI & Cieplak M. *Chirality and protein folding*. Journal of Physics: Condensed Matter 2005. 17(18):S1565.

[112] Tsai J, Taylor R, Chothia C & Gerstein M. *The packing density in proteins: standard radii and volumes*. Journal of Molecular Biology 1999. 290(1):253–266.

[113] Cieplak M, Hoang TX & Robbins MO. *Thermal effects in stretching of Go-like models of titin and secondary structures*. Proteins: Structure, Function, and Bioinformatics 2004. 56(2):285–297.

[114] Hess B, Kutzner C, van der Spoel D & Lindahl E. *GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation*. Journal of Chemical Theory and Computation 2008. 4(3):435–447. doi:10.1021/ct700301q.

[115] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW & Kollman PA. *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*. Journal of the American Chemical Society 1995. 117(19):5179–5197.

[116] Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen My, Pieper U & Sali A. *Comparative protein structure modeling using modeller*. John Wiley & Sons, Inc. ISBN 9780471250951, 2002. doi: 10.1002/0471250953.bi0506s15.

[117] Qiu D, Shenkin PS, Hollinger FP & Still WC. *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii*. The Journal of Physical Chemistry A 1997. 101(16):3005–3014.

[118] Bussi G, Donadio D & Parrinello M. *Canonical sampling through velocity rescaling*. The Journal of Chemical Physics 2007. 126(1):014 101.

[119] Nosé S. *A molecular dynamics method for simulations in the canonical ensemble*. Molecular Physics 1984. 52(2):255–268.

[120] Hoover WG. *Canonical dynamics: equilibrium phase-space distributions*. Physical Review A 1985. 31(3):1695.

[121] Ryckaert JP, Ciccotti G & Berendsen HJ. *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*. Journal of Computational Physics 1977. 23(3):327–341.

[122] Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, Raiteri P, Donadio D, Marinelli F, Pietrucci F, Broglia RA & Parrinello M. *PLUMED: A portable plugin for free-energy calculations with molecular dynamics*. Computer Physics Communications 2009. 180(10):1961 – 1972. ISSN 0010-4655. doi:http://dx.doi.org/10.1016/j.cpc.2009.05.011.

[123] Cossio P, Trovato A, Pietrucci F, Seno F, Maritan A & Laio A. *Exploring the universe of protein structures beyond the Protein Data Bank.* PLoS Computational Biology 2010. 6(11):e1000 957. doi:10.1371/journal.pcbi.1000957.

[124] Oroz J, Hervás R, Valbuena A & Carrión-Vázquez M. *Unequivocal single-molecule force spectroscopy of intrinsically disordered proteins*. Methods in Molecular Biology 2012. 896:71–87. doi:10.1007/978-1-4614-3704-8_5.

[125] Hervás R, Fernández-Ramírez M, Abelleira L, Laurents D & Carrión-Vázquez M. *Nanomechanics of neurotoxic proteins. Insights at the start of the neurodegeneration cascade*. 2013. doi:10.1016/B978-0-12-394431-3.00006-7.

[126] Sikora M, Sułkowska JI, Witkowski BS & Cieplak M. *BSDB: the biomolecule stretching database*. Nucleic Acids Research 2011. 39(suppl 1):D443–D450.

[127] Wallin S, Zeldovich KB & Shakhnovich EI. *The folding mechanics of a knotted protein*. Journal of Molecular Biology 2007. 368:884–893.

[128] Chwastyk M & Cieplak M. *Cotranslational folding of deeply knotted proteins*. Journal of Physics: Condensed Matter 2015.

[129] Chwastyk M, Poma AB & Cieplak M. *Statistical radii associated with amino acids to determine the contact map: fixing the structure of a type I cohesin domain in the Clostridium thermocellum cellulosome*. Physical Biology 2015.

[130] Cieplak M & Sułkowska JI. *Structure-based models of biomolecules: stretching of proteins, dynamics of knots, hydrodynamic effects, and indentation of virus capsids*. In A Kolinski, editor, *Multiscale approaches to protein modeling: structure prediction, dynamics, thermodynamics and macromolecular assemblies*, pages 179–208. Springer, 2010.

[131] Noel JK, Whitford PC & Onuchic JN. *The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function*. Journal of Physical Chemistry B 2012. 116:8692–8702.

[132] Sobolev V, Wade RC, Vriend G & Edelman M. *Molecular docking using surface complementarity*. Proteins: Structure, Function, and Bioinformatics 1996. 25(1):120–129.

[133] González Á. *Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices*. Mathematical Geosciences 2010. 42(1):49–64.

[134] Chwastyk M, Jaskólski M & Cieplak M. *Structure-based thermodynamic and mechanical stability of plant PR-10 proteins with cavities* . FEBS Journal 2014. 281:416–429.

[135] Zahn R, Liu A, Lührs T, Riek R, von Schroetter C, García FL, Billeter M, Calzolai L, Wider G & Wüthrich K. *NMR solution structure of the human prion protein*. Proceedings of the National Academy of Sciences 2000. 97(1):145–150.

[136] Danielsson J, Andersson A, Jarvet J & Gräslund A. *15N relaxation study of the amyloid β-peptide: structural propensities and persistence length*. Magnetic Resonance in Chemistry 2006. 44(S1):S114–S121.

[137] Mukrasch MD, Bibow S, Korukottu J, Jeganathan S, Biernat J, Griesinger C, Mandelkow E & Zweckstetter M. *Structural polymorphism of 441-residue tau at single residue resolution*. PLoS Biology 2009. 7(2):e1000 034.

[138] Sandal M, Valle F, Tessari I, Mammi S, Bergantino E, Musiani F, Brucale M, Bubacco L & Samorì B. *Conformational equilibria in monomeric α-synuclein at the single-molecule level*. PLoS Biology 2008. 6(1):e6.

[139] Brucale M, Sandal M, Di Maio S, Rampioni A, Tessari I, Tosatto L, Bisaglia M, Bubacco L & Samorì B. *Pathogenic mutations shift the equilibria of α-synuclein single molecules towards structured conformers*. ChemBioChem 2009. 10(1):176–183.

[140] Dougan L, Li J, Badilla CL, Berne B & Fernandez JM. *Single homopolypeptide chains collapse into mechanically rigid conformations*. Proceedings of the National Academy of Sciences 2009. 106(31):12 605–12 610.

[141] Kelly JW. *The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways*. Current Opinion in Structural Biology 1998. 8(1):101–106.

[142] Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C & Vriend G. *A series of PDB related databases for everyday needs.* Nucleic Acids Res 2011. 39(Database issue):D411–D419. doi:10.1093/nar/gkq1105.

[143] Zhang Y & Skolnick J. *Scoring function for automated assessment of protein structure template quality*. Proteins: Structure, Function, and Bioinformatics 2004. 57(4):702–710. ISSN 1097-0134. doi:10.1002/prot.20264.

[144] Zhang Y & Skolnick J. *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res 2005. 33(7):2302–2309. doi:10.1093/nar/gki524.

[145] Cieplak M, Allan DB, Leheny RL & Reich DH. *Proteins at air-water interfaces: A coarse-grained model*. Langmuir 2014. 30(43):12 888. doi:10.1021/la502465m. PMID: 25310625.

[146] Sikora M, Szymczak P, Thompson D & Cieplak M. *Linker-mediated assembly of gold nanoparticles into multimeric motifs*. Nanotechnology 2011. 22(44):445 601.

[147] Maxwell JC. *L. on the calculation of the equilibrium and stiffness of frames*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Sciences 1864. 27(182):294–299.

[148] Carrion-Vazquez M, Oberhauser AF, Fowler SB, Marszalek PE, Broedel SE, Clarke J & Fernandez JM. *Mechanical and chemical unfolding of a single protein: a comparison*. Proceedings of the National Academy of Sciences USA 1999. 96(7):3694–3699.

[149] Sułkowska JI, Sułkowski P, Szymczak P & Cieplak M. *Tightening of knots in proteins*. Physical Review Letters 2008. 100(5):058 106.

[150] Virnau P, Mirny LA & Kardar M. *Intricate knots in proteins: Function and evolution*. PLoS Computational Biology 2006. 2(9):e122.

[151] Krishnan R, Goodman JL, Mukhopadhyay S, Pacheco CD, Lemke EA, Deniz AA & Lindquist S. *Conserved features of intermediates in amyloid assembly determine their benign or toxic states*. Proceedings of the National Academy of Sciences 2012. 109(28):11 172–11 177.

[152] Hervás R, Majumdar A, Li L, del Carmen Fernández-Ramírez M, Unruh J, Slaughter B, Galera-Prat A, Santana E, Suzuki M, Nagai Y, Bruix M, Casas-Tintó S, Menéndez M, Laurents DV, Si K & Carrión-Vázquez M. *Structural basis of memory consolidation mediated by Orb2 amyloid*. PLoS Biology In press.

[153] Zhang F, Hu M, Tian G, Zhang P, Finley D, Jeffrey PD & Shi Y. *Structural insights into the regulatory particle of the proteasome from Methanocaldococcus jannaschii*. Molecular Cell 2009. 34(4):473–484.

[154] Groll M, Ditzel L, Löwe J, Stock D, Bochtler M, Bartunik HD & Huber R. *Structure of 20S proteasome from yeast at 2.4 Å resolution.* Nature 1997. (386):463–71.

[155] Cras J, Rowe-Taitt C, Nivens D & Ligler F. *Comparison of chemical cleaning methods of glass in preparation for silanization*. Biosensors and Bioelectronics 1999. 14(8):683–688.

[156] Halliwell CM & Cass AE. *A factorial analysis of silanization conditions for the immobilization of oligonucleotides on glass surfaces*. Analytical Chemistry 2001. 73(11):2476–2483.

[157] Burnham NA, Chen X, Hodges CS, Matei GA, Thoreson EJ, Roberts CJ, Davies MC & Tendler SJB. *Comparison of calibration methods for atomic-force microscopy cantilevers*. Nanotechnology 2003. 14(1):1.

[158] Hutter JL. *Comment on tilt of atomic force microscope cantilevers: effect on spring constant and adhesion measurements*. Langmuir 2005. 21(6):2630–2632. doi:10.1021/la047670t. PMID: 15752063.

Part V

Appendices

A. SMFS experiment protocol

In this appendix I give step-by-step protocols for the sample preparation and the handling of the AFM for SMFS as used in this work, as well as some considerations about the experimental procedures.

A.1. Coverslip functionalization protocol

A.1.1. Materials

- 60 circular glass coverslips (15 mm of diameter)
- 100 mL 1:1 MeOH in HCl:
  - 50 mL Metanol (MeOH)
  - 50 mL Hydrochloric acid (HCl)
- 4 $\times$ 100 mL milliQ water
- (2 + 2) $\times$ 100 mL absolute ethanol
- (1 + 1) $\times$ 100 mL toluene

- 100 mL 5 % (v/v) 3MPTS in toluene:
  - 5 mL 3(mercaptopropyl)trimetiloxylane (3MPTS)
  - 95 mL toluene
- 45 mL 50 mM Tris pH 8.4.
  - 2.25 mL 1 M Tris (formerly prepared).
  - 42.75 mL milliQ water.
  - HCl and NaOH to adjust.
- 50 mL 100 mM DTT in 50 mM Tris pH 8.4.
  - 0.77 g dithiothreitol (DTT).
  - 45 mL 50 mM Tris.
  - HCl and NaOH to adjust.

A.1.2. Procedure – Day 1

1. First wash: 1:1 MeOH + HCl [155].
   - Prepare 100 mL of a 1:1 solution of MeOH and HCl.
   - Pour the solution carefully into a Petri dish.
   - Put the glass coverslips in the Petri dish with the solution.
   - Cover and gently agitate the Petri dish.
   - Wait for 30 minutes.
2. Second wash: MilliQ water.
   - Pour 100 mL of MilliQ water carefully into a new Petri dish (or the cover of the one we were using).
   - Completely remove the liquid in the Petri dish containing the glass coverslips. Using a Pasteur pipette, make sure all the remaining liquid has been removed.
   - Move the glass coverslips one by one from the (now) dry Petri dish to the one with clean MilliQ water.

- Cover and gently agitate the Petri dish.
- Repeat 3 more times (total of 4).

3. Third wash: Absolute ethanol.

   - Pour 100 mL of absolute ethanol carefully into a new Petri dish (or the cover of the one we were using).
   - Completely remove the liquid in the Petri dish containing the glass coverslips. Using a Pasteur pipette, make sure all the remaining liquid has been removed.
   - Move the glass coverslips one by one from the (now) dry Petri dish to the one with clean ethanol.
   - Cover and gently agitate the Petri dish.
   - Repeat 1 more times (total of 2).

4. Fourth wash: Toluene.

   - Pour 100 mL of toluene carefully into a new Petri dish (or the cover of the one we were using).
   - Completely remove the liquid in the Petri dish containing the glass coverslips. Using a Pasteur pipette, make sure all the remaining liquid has been removed.
   - IMPORTANT: Before this steps, the coverslips must be completely dry. If needed, wait for a few minutes for the remnant of the ethanol to evaporate.
   - Move the glass coverslips one by one from the (now) dry Petri dish to the one with toluene.
   - Cover and gently agitate the Petri dish.

5. 3MPTS deposition [156].

   - Prepare 100 mL of 5 % (v/v) solution of 3MPTS in toluene (5 mL of 3MPTS in 95 mL of toluene).
   - Pour the solution carefully into a new Petri dish (or the cover of the one we were using).

- Completely remove the liquid in the Petri dish containing the glass coverslips. Using a Pasteur pipette, make sure all the remaining liquid has been removed.
- Move the glass coverslips one by one from the (now) dry Petri dish to the one with the solution.
- Cover the Petri dish and wrap it completely with aluminium foil.
- Place the Petri dish on an orbital shaker at 30 rpm during 4 hours.

6. Fifth wash: Repeat toluene wash (step 4).
7. Sixth wash: Repeat absolute ethanol wash (step 3).
8. Cure in an oven over night.

A.1.3. Procedure – Day 2

9. Reduction of unwanted S–S bonds.
   - Prepare 50 mL of 100 mM DTT in 50 mM Tris at pH 8.4 (adjust with HCl and NaOH).
   - Pour the solution carefully into a new Petri dish (or the cover of the one we were using).
   - Completely remove the liquid in the Petri dish containing the glass coverslips. Using a Pasteur pipette, make sure all the remaining liquid has been removed.
   - Move the glass coverslips one by one from the (now) dry Petri dish to the one with the solution.
   - Cover and gently agitate the Petri dish.
   - Wait for 15 minutes.
10. Seventh wash: Repeat absolute ethanol wash (step 2, note the 4 times repetition).
11. NTA-Ni deposition.

- Completely remove the liquid in the Petri dish containing the glass coverslips. Using a Pasteur pipette, make sure all the remaining liquid has been removed.
- Adjust, if needed, the pH of 10 mM MOPS solution to 7.0 using KOH and HCl.
- Pour 650 µL of MOPS solution directly into the NTA-maleimido test tube.
- Pair up the (now dry) coverslips.
- Deposit 19 µL of the resulting solution on one of the coverslips of each pair, then sandwich it with its couple.
- Wait for 30 minutes.
- IMPORTANT: From now on, orientation is important, since only one of the faces of the coverslip is functionalized.
- Carefully separate each couple. After the separation, wash each coverslip briefly in milliQ water and dry on blotting paper. Make sure to leave it with the functionalized side up and to keep the pairing.
- Deposit 19 µL of $NiCl_2$ on one of the coverslips of the couple and sandwich with the other.
- Wait for 10 minutes.
- Carefully separate each couple. After the separation, wash each coverslip briefly in milliQ water and dry on blotting paper. Make sure to leave it with the functionalized side up.

A.2. AFM experiment preparation

A.2.1. First Steps

- To book the AFM for a working day, write your name on the AFM calendar on-line.
- If you notice there is an experiment already running, ask before you do anything, then save the work and close the experiment properly (see sec. A.2.4) before you start with your own.

A.2.2. Setting up a SMFS Experiment

NOTE: Remember you should always use latex gloves for manipulating lab substances or materials.

Preparing your substrate

1. Take an old metallic disk and remove the coverslip and the two sided sellotape[1] using a scalpel. Make sure it is clean for reuse.

2. Using two sided sellotape, glue your coverslip to the metallic disk. Remember to avoid as much as possible touching the functionalized surface in order not to remove its properties, or add undesired residues.

3. Remove the head of the AFM from its position and put it carefully behind the magnets. Be careful of the magnetic forces.

4. Place your metallic disk on the magnets approaching them from the side.

Opt Replace the head of the AFM to place, then turn on the laser and see the place the dot is on, so that you can place the protein directly under it. Then, remove the head again.

Sample incubation

1. Using a pipette, take the desired amount of protein and buffer (less than 50 μL), and put them on the coverslip.

2. Incubate the protein for 10 to 30 minutes (depending on substrate and protein concentration).

Preparing the fluid cell

1. While you wait for the incubation to be over, use the microscope or the magnifying glass to look for a chip with at least one of the small cantilevers.

[1] If you cannot remove completely the sellotape, you may want to try isopropanol in order to soften it. However, it becomes fluid and very sticky.

2. Use the pincers to grab it and deposit it in the UV/Ozone cleaner for about 10 s. Remember to put the tips to the right.

3. Make working buffer and fill two syringes with it. Then place them in the holes of the fluid cell.

4. Using the pincers again, take the chip out of the UV/Ozone cleaner and put it directly in the fluid cell. Once placed, it has to be tilted, since we put it on a ramp, and well fixed in place by the spring.

5. Using the syringes, fill the O-ring of the fluid cell with buffer. Then turn it downwards and keep injecting buffer into the cell in order to clean it, until there is about 0.2 to 0.5 mL in each syringe.

Finding a tip

1. Before you put the fluid cell in the AFM head, you need to start the electronics in the correct order:

   - Plug the laser to the current stabilizer.
   - Turn it on using the key.
   - Turn on the oscilloscope.
   - Turn on the JRC electronics using the red switch at the front.
   - Turn on Dulcinea using the switch at the back.
   - Do not turn on the PI electronics.

2. It is advisable to retire some buffer from the fluid cell in order to avoid overflows. Place the fluid cell face down in the cavity of the AFM head, first tilted forward and then take the back down on to the surface.

3. Fix the cell to the base of the cavity using the appropriate screw.

4. Place a piece of paper on one of the fixing legs and check the diffraction patterns on it. Fig. A.1 shows possible designs you can see on the paper.

5. Once you have found your tip, remove the piece of paper.

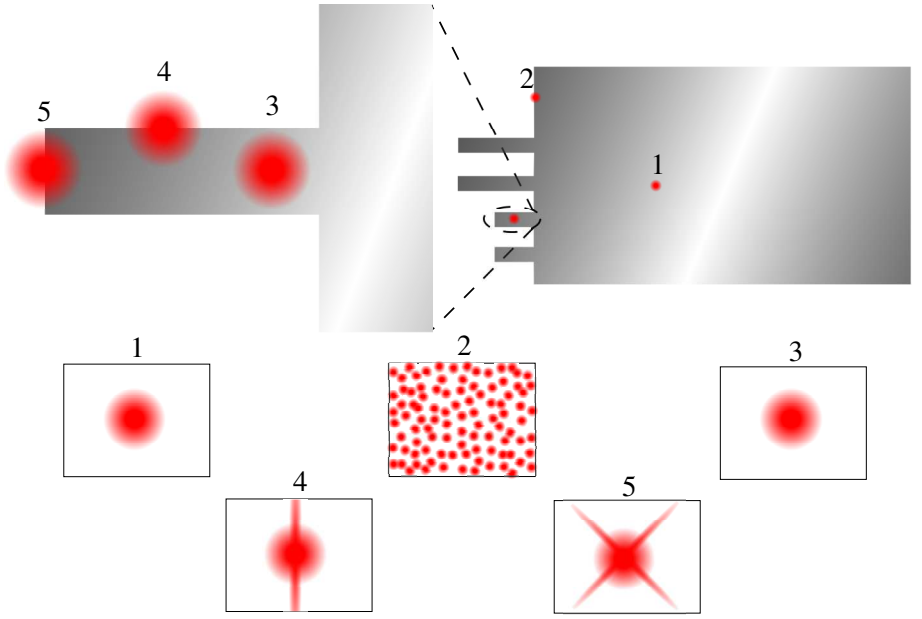


Figure A.1: Guide for focusing the laser beam on the cantilever. Each panel shows the diffraction pattern that can be seen while locating the tip and where is the spot when it can be seen. Note that 1 and 3 are quite similar and can be confused.

6. Using the mirror lever, increase the $A+B$ value of the JRC electronics to the maximum value, and try to put the $A-B/A+B$ one as near to zero as possible.

7. If really needed, using the appropriate screw to put the $A-B/A+B$ value to zero. If you notice it has a tendency to increase or to decrease, put the value a bit lower or higher, respectively.

Cantilever Calibration

1. Open *Igor Pro* (it takes some time to load), then go to *File* > *Open File* > *Procedure...*, then load the file in the path *C:\Archivos de Programa\WaveMetrics\Igor Pro Folder\User Procedures\SMFSiAFM_Igor6.1.ipf*.

When the procedure loads, click on *Compile* at the bottom of the window, and minimize it.

2. Go to the menu *Macros* > *Initialize DAQ*. This will make a new window appear, named *The AFM*.

3. On the *Calibrate* tab, click the button that says *Power* to display the power spectrum. Check that the resonance peak corresponds to the frequency of your tip and place the cursors at the minima around it, then click *Power* again.

4. Turn on the PI controller using the switch at the back.

5. Open WsXM 4.0 Develop 11.4.

6. Click on the button labelled *DA*, and then on *GO*. Wait for the windows to load and:

   - Rise the XY gain to 15 and click *Update*.
   - Turn on the XY and Z external scan controls, the *Z closed loop scanner* and the *XY linear feedback*.
   - Open the motor control panel.

7. Check the $A-B/A+B$ value (normal force from now on) and make it zero with the lever if necessary. Then click on *Approach* under the *Move* mode (left) in the Motor panel and wait until you become in range[2].

8. Once the message *In range* appears on your screen, before clicking on *OK*, make sure that the normal force is still zero. If it is not, use the *HeadNull* button on *The AFM* window. Then click on *OK* in the WsXM window and check the following:

   - The green bar in the motor panel in WsXM must be a bit over the middle. If it goes up to the top, we need to approach again.
   - At the *Main Linear Feedback* subsection in WsXM, the *Z (nm)* value must be around 350. If it is higher, we must approach again.

[2]Should a Warning prompt, telling you that the laser is going to be turned on, simply click on yes, since the laser does not depend on this software (we turned it on before already).

9. Once we are in range, we disable the Main Feedback[3], we change the value of *Z(nm)* to 0 and click on the *Slope* button on *The AFM* window. This rises the surface so that the cantilever presses on it and we get a first $F(z)$ curve, which we use to determine the elastic constant of the probe. On *The AFM* panel, the values *SC (pN/nm)* and *Slope (nm/V)* are computed, so we can check if they are the ones they should be (close to the manufacturer's specification). If they are, we are ready to start with the data acquisition. If they are not, we need to make it as good as possible by either:

    - Moving through the plane with the *X pos* and *Y pos* bars in *The AFM* window and check again.
    - Moving far from the surface with the *Z pos* bar and increasing the *amplitude (nm)* value (pushing harder). Be careful not to press too hard, or you might break your tip.

Starting Data Acquisition

1. Once we have the tip parameters calibrated, we click on *FX* on *The AFM* panel. Here we must change the parameters according to our sample:

    - *amplitude (nm)* controls the jumping distance, so it must be a bit larger than the full length of the protein.
    - *size* and *# Points* control (as their names indicate) the size we are going to sweep and the points we will have in this size. These values are to be large enough if we want the experiment to run overnight, but since the drift can also make the experiment stop, it is advisable to have a stable system in order to scan a big area.

2. Once you have changed the parameters to suit your own experiment, we need to save it with a proper name using *File* > *Save Experiment As...*, then click on *Go* in *The AFM* window to start the acquisition.

3. While the experiment runs, due to the drift, the *Z pos* value might move. If this is the case, every time it is too far from the origin, you will need to stop the experiment holding the *Ctrl* button and:

[3]There may be another warning here, telling us that we might crash the tip if we disable the Feedback. Simply ignore it again and click *Yes*.

- Click on the *Zero* button on *The AFM* window and then use the lever at the back of the head of the AFM to bring the normal force to zero.
- If the deviation is positive, simply enter 0 in the *Z pos* value, and the PI will retract.
- If the value is negative we must first move the piezoelectric away since it would make the tip crash. To do that, we click on *Withdraw* in the motor panel of WsXM preferably the one under *Steps* (right) at least until the *Z pos* value under the *Main Linear Feedback* window in WsXM is larger than the value you set for the amplitude. Then enter 0 in the *Z pos* value and the PI will go back to place. Be careful not to enter 0 before withdrawing, or the tip is likely to crash.

4. If you intend to leave the AFM unattended, it is a good idea to use the autodetect mode:

   - Not autodetect mode: All trials will be stored, so you will have a lot of curves in very little time, most of which will not be useful in the experiment. This can generate a memory overflow, so you have to check on your experiment every now and then and avoid the number of saved curves to be over 1000.
   - Pattern autodetect mode: By checking on the first check box (*detect pattern*), only curves that have more than one peak will be stored. The number of stored curves will dramatically reduce, and some valid curves might not be taken, but this way you can leave the experiment unattended for a longer period.
   - Force autodetect mode: By checking both check boxes (*detect pattern* and *detect force*), only the curves with higher force than specified in the *min force (pN)* box will be registered. This one also reduces a lot the number of stored curves, but saves more than the previous method.

A.2.3. First sieve

After a while you need to delete the curves that are not useful. Some of those curves are tricky to see, but others are easily identified:

- Curves with no peaks,

- Curves with just one peak,
- Curves that keep rising (which means something got attached to the tip and we did not pull enough to detach it),
- Curves that are obviously too long,
- Curves that have more markers than they should,
- etc.

However, this depends a lot on the molecule under study, so these general premises should not be taken as rules.

1. *Igor Pro* cannot keep up with the data acquisition and simultaneously erase or analyze curves. Thus, if the experiment is still running and you want to erase curves, stop acquisition by keeping the *Ctrl* key pressed for a few seconds[4]. When the acquisition stops, the bottom left corner will change from *Abort* to *Ready*.

2. To erase curves, you need to load an analysis procedure. To do that, go to *File* > *Open File* > *Procedure...*, then load one of the files in the path *C:\Archivos de Programa\WaveMetrics\Igor Pro Folder\User Procedures* that is used for analysis[5]. When the procedure loads, click on *Compile* at the bottom of the window, and minimize it.

3. Go to the menu *Macros* > *Analysis*. This will make a new window appear, named *AFM Analysis*.

4. To visualize some curves, edit the fields *start#* (the curve number you want to start with) and *display#* (the number of curves you want to display each time) under *Display*. When you fill both, the curves will appear.

5. To erase some curves, under *FX Delete*, fill the boxes *From:* and *To:* with the first and the last curves of the range we want to delete. The curves corresponding to the numbers you enter in the *From:* and *To:* boxes will also be deleted. Then press the *Delete!* button.

[4] Make sure that, while you press the *Ctrl* key, *Igor Pro* window has focus.

[5] At the moment they are *Force_Analysis_ceroA.ipf*, *SMFSiAFM_Analysis1.0.ipf* and *SMFSiAFM_Analysis1.1.ipf*; I use *SMFSiAFM_Analysis1.1.ipf* in the examples

6. Repeat this operation until you have deleted the curves you don't want. Then save the experiment by going to *File* > *Save Experiment*.

7. After this, you can restart your experiment if you want to, or end the experiment.

A.2.4. Ending the experiment

To end your experiment and being able to take it to other computers, the following steps are needed[6].

1. If the data acquisition is active (if the bottom-left corner of *Igor Pro* window says *Abort*) make sure that *Igor Pro* window is focused and keep the *Ctrl* key pressed for a few seconds (until the bottom-left corner changes to *Ready*).

2. Click on the *Withdraw* button under the *Move* mode (left) in the *WSxM* motor panel, wait for 10 to 15 s, then click on the same button (it now says *Stop*).

3. Now that we are far it is safe to stop the signals from *Igor Pro*. In *The AFM* panel, write zero in the three boxes at the top (*Z pos*, *X pos*, *Y pos*), then press *Zero* and *Reset* buttons.

4. Under *Windows* > *Procedure windows*, click on the name of the acquisition procedure you loaded in subsection A.2.2 and a window with the code will pop up. Close the window and, when prompted, choose *Kill*.

5. It is now safe to save the experiment under *File* > *Save Experiment*, and close *Igor Pro*.

6. Eventually, to close *WSxM*, disable the XY and Z external scan controls, the *Z closed loop scanner* and the *XY linear feedback*. Then press stop button and, when the main area in the window goes back to gray, close the window.

7. As a last step, turn the PC off. Also, turn off the laser using the key and disconnect it from the insulator.

[6] Actually they are not necessary, but extremely highly recommendable

A.3. AFM tip calibration

When buying a cantilever wafer, many factors are at play that may affect the elastic constant of the cantilever. Therefore, even if the manufacturer gives an approximation of the expected elastic constant, it is important to do a calibration of any single cantilever one uses in an experiment. There are several ways to calibrate an AFM tip [157], of which the one used in the laboratory is the Thermal Fluctuations one[53], which proceeds in two steps:

A.3.1. Sensitivity

The sensitivity is not so much a property of the cantilever but of its interaction with its environment. It relates the changes in voltage between the two halves of the photodiode with actual movement (measured as a distance). This relation depends on many factors, such as the refractive index of the buffering solution the cantilever is submerged in, the reflectivity of the cantilever itself or the light intensity received by the photodiode, among others. A good way to measure this property consists on moving the cantilever a known distance in a controlled fashion, and comparing the change in voltage detection in the photodiode to the known displacement.

In practice, this is accomplished by pushing a hard, non-deflectable surface against the cantilever tip using a constant velocity movement of the piezoelectric positioner. The graph is then divided in four regions: free-approach, contact-approach, contact-retreat and free-retreat. The free zones are characterized by a constant value of the photodiode-detected intensity as the positioner moves, while in the contact zones the registered voltage and the displacement length present a linear relationship between one-another. The transition between free and contact regions are typically smooth, although the exact shape depends on the interaction of the tip and the surface.

Once the curve has been taken, a straight line is fitted to the contact regions, the slope of which yields the sensitivity, in voltage over distance, of the cantilever.

One might think that the sensitivity might be different when pushing (as it is measured) and when pulling (as when one does the experiment with proteins and measures the force). This, however, is not the case, as can be assessed by measuring the sensityvity of the cantilever on a mica substrate in air (*i.e.* with no buffering solution). This conditions maximize the attractive interaction of the tip

and the substrate, leading to a longer adhesion and thus to the ability to compare the pushing and pulling sensitivities.

A.3.2. Spring constant determination

The spring constant, in contrast with the sensitivity, is only a property of the cantilever. To access this value, one needs to check the movement of the cantilever due to Brownian motion. In a first order approximation, the cantilever can be considered a one-dimensional elastic spring of constant k, which must be obtained using the equipartition theorem.

In a one-dimensional system, the thermal energy is given by equation A.1. In the case of the system being an elastic spring, this energy is completely transformed into elastic energy, equation A.2. The combination of the two leads to equation A.3, which gives us the spring constant of the cantilever.

$$E_{1D} = \frac{1}{2} k_B T \tag{A.1}$$

$$E_{spring} = \frac{1}{2} k \langle x^2 \rangle \tag{A.2}$$

$$k = \frac{k_B T}{\langle x^2 \rangle} \tag{A.3}$$

However, $\langle x^2 \rangle$ cannot be directly measured in the system, but is obtained as the fluctuation of the laser intensity difference in the photodiode. Therefore, the previously measured sensitivity is needed in order to compute the spring constant in standard units of force over distance.

Nonetheless, the system is not a one-dimensional spring, and therefore the fluctuations in the signal can be due to many other factors, such as torsion of the cantilever or horizontal vibration. To get rid of this, one studies the frequency response and eliminates all excited frequencies that do not correspond to the main vibration mode.

After the removal of the undesired peaks in the Fourier-transformed intensity signal, integrating over all possible frequencies yields the expected $\langle V^2 \rangle$, which can in turn be converted to distance using the sensitivity.

Several years after the publication of this method, the same author commented [158] that it is not enough to remove spurious resonance peaks to measure the correct spring constant, but some corrections are needed based on several

incorrect assumptions, such as the cantilever being a one-dimensional spring, the fact that one end of the cantilever cannot move or the torsion of the cantilever. This discovery yielded several corrections to the typically used formula for the spring constant, which are predicted to be small [157] and are therefore not commonly applied.

B. Statistical Analysis Code

The error bars used in the figures in this work show a 95 % confidence interval for the data based on the experimental or theoretical results. The calculations of the statistics of this interval were carried out using home-made Python scripts that apply the smoothed bootstrapping method. These scripts use the `numpy` package (imported as `np`) and are detailed next.

- Firstly, a Cumulative Density Function (CDF) is computed from the data. This is done by generating a list of x values (bins), and counting how many elements in the data list are below each of those values. The binning is performed manually because in this way it can be done logarithmically as well as linearly (or with any other desired shape). The CDF is eventually normalized to 1.

```
def make_cdf(cdf_x, data):
    '''
    Computes the CDF of a list of data.
    start is the smallest value, binwidth is the size of
    each bin and nbins is the number of points the CDF is
    to have.
    Returns a list with nbins points, where the value at
```

```
    index x is the probability of finding a value in data
    that is smaller than the corresponding value in cdf_x
    '''
    length = float(len(data))
    cdf = [len([y for y in data if y <= x])/length \
            for x in cdf_x]
    return cdf
```

- Next, a Probability Density Function (PDF)[1] can be computed by taking the derivative of the CDF.

```
def make_pdf(cdf_x, cdf):
    '''
    Computes the PDF by taking the derivative of the CDF
    Includes the first element in the second and makes the
    first zero.
    '''
    dx = np.gradient(cdf_x)
    pdf = list(np.gradient(cdf,dx))
    pdf[1] += pdf[0]
    pdf[0] = 0.
    return pdf
```

- In order to resample a distribution, the inverse of the CDF needs to be computed. In this script we use the following `find_level` function, which finds the two points between which the value lies and makes a linear approximation between them.

```
def find_level(fx, fy, y):
    '''
    Finds the rightmost x value where fy[x]=y
    '''
    assert len(fx) == len(fy), \
        "Both lists must be of the same length"
    fx = sorted(fx)
    if y > max(fy):
```

[1]For discrete sampling, this is normally called Probability Mass Function or PMF, but since our computation is *pseudo*-analytical in that it comes from differentiating the CDF we decided to stick to the continuous notation.

```
        return max(fx)
    k = len(fy)-1
    try:
        while fy[k] > y:
            k -= 1
    except IndexError:
        return 0.
    return fx[k]+(y-fy[k])*(fx[k+1]-fx[k])/(fy[k+1]-fy[k])
```

- The bootstrapping function resamples the CDF to generate new simulated experiments. From these, it computes new CDFs and PDFs. The resampling is done taking uniformly distributed random numbers in $[0,1)$, calculating the inverse image of these numbers through the (original) CDF and adding to this number a small normally distributed noise with a variance equal to $1/\sqrt{N}$, where N is the sample size. The addition to this term comes from the smoothed bootstrapping method, which is better to reduce the effect of outliers than common bootstrapping. This function generates many distributions, which are stored to be treated as desired further on.

```
def bootstrap(cdf_x, cdf, rsnum, bsnum):
    '''
    Computes bsnum CDFs and PDFs from one original CDF
    Does bsnum bootstrap resamples with probability cdf,
    each resample having rsnum points.
    '''
    pdfs = []
    cdfs = []
    while bsnum > 0:
        bsnum -= 1
        gauss_var = 1./math.sqrt(rsnum)
        new_data = [find_level(cdf_x, cdf, \
            random.random()) + \
            abs(random.gauss(0, gauss_var)) \
            for i in range(rsnum)]
        new_cdf = make_cdf(cdf_x , new_data)
        cdfs.append(new_cdf)
        pdfs.append(make_pdf(new_cdf, cdf_x))
    return pdfs, cdfs
```

- Finally, the script implements a function to find the median, as well as the higher and lower values to determine the 95 % confidence interval.

```
def stats(function_list):
    '''
    Computes median and high and low ends of a 95 %
    confidence interval given a list of CDF or PDF
    functions. The input must be a list of lists, so
    that each inner list is a complete CDF or PDF.
    If it were a matrix, each row would be a PDF and
    each column one value of x.
    '''
    med = list(np.percentile(function_list, \
        50, axis=0))
    high = list(np.percentile(function_list, \
        97.5, axis=0))
    low = list(np.percentile(function_list, \
        2.5, axis=0))
    return med, high, low
```

List of publications

1. Galera-Prat A, **Gómez-Sicilia À**, Oberhauser AF, Cieplak M & Carrión-Vázquez M. *Understanding biology by stretching proteins: recent progress.* Current Opinion in Structural Biology 2010. 20(1):63–69.

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Understanding biology by stretching proteins: recent progress

Albert Galera-Prat[1,2,6], Angel Gómez-Sicilia[1,2,6], Andres F Oberhauser[4], Marek Cieplak[5] and Mariano Carrión-Vázquez[1,2,3]

Single molecule manipulation techniques combined with molecular dynamics simulations and protein engineering have enabled, during the last decade, the mechanical properties of proteins to be studied directly, thereby giving birth to the field of protein nanomechanics. Recent data obtained from such techniques have helped gain insight into the structural bases of protein resistance against forced unfolding, as well as revealing structural motifs involved in mechanical stability. Also, important technical developments have provided new perspectives into protein folding. Eventually, new and exciting data have shown that mechanical properties are key factors in cell signaling and pathologies, and have been used to rationally tune these properties in a variety of proteins.

enabled single molecules to be studied, avoiding ensemble averaging. Among other adventages, these techniques can capture transient intermediates and alternative conformers. In particular, the techniques used to manipulate molecules individually include atomic force microscopy (AFM), optical tweezers, and magnetic tweezers [1].

An important feature of single molecule experiments is that they are closely comparable to molecular dynamics (MD) simulations of individual molecules. MD simulations provide an atomic description of the system, not accessible experimentally. They have been proved to be very accurate, and even predictive, such that they offer an

2. Hervás R, Oroz J, Galera-Prat A, Goñi O, Valbuena A, Vera AM, **Gómez-Sicilia À**, Losada-Urzáiz F, Uversky VN, Menéndez M, Laurents DV, Bruix M & Carrión-Vázquez M. *Common features at the start of the neurodegeneration cascade.* PLoS Biology 2012. 10(5):1014.

OPEN ACCESS Freely available online PLoS BIOLOGY

Common Features at the Start of the Neurodegeneration Cascade

Rubén Hervás[1,2☯], Javier Oroz[1,2☯], Albert Galera-Prat[1,2], Oscar Goñi[1,2], Alejandro Valbuena[1,2], Andrés M. Vera[1,2], Àngel Gómez-Sicilia[1,2], Fernando Losada-Urzáiz[1,2], Vladimir N. Uversky[3,4], Margarita Menéndez[5], Douglas V. Laurents[6], Marta Bruix[6], Mariano Carrión-Vázquez[1,2]*

1 Instituto Cajal, IC-CSIC & Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain, **2** Instituto Madrileño de Estudios Avanzados en Nanociencia (IMDEA-Nanociencia), Madrid, Spain, **3** University of South Florida, College of Medicine and Byrd Alzheimer's Research Institute, Tampa, Florida, United States of America, **4** Institute for Biological Instrumentation. Russian Academy of Sciences, Pushchino, Moscow Region, Russia, **5** Instituto de Química-Física Rocasolano, IQFR-CSIC & Centro de Investigación Biomédica en Red sobre Enfermedades Respiratorias (CIBERES), Madrid, Spain, **6** Instituto de Química-Física Rocasolano, IQFR-CSIC, Madrid, Spain

Abstract

Amyloidogenic neurodegenerative diseases are incurable conditions with high social impact that are typically caused by specific, largely disordered proteins. However, the underlying molecular mechanism remains elusive to established techniques. A favored hypothesis postulates that a critical conformational change in the monomer (an ideal therapeutic target) in these "neurotoxic proteins" triggers the pathogenic cascade. We use force spectroscopy and a novel methodology for unequivocal single-molecule identification to demonstrate a rich conformational polymorphism in the monomer of four representative neurotoxic proteins. This polymorphism strongly correlates with amyloidogenesis and neurotoxicity: it is absent in a fibrillization-incompetent mutant, favored by familial-disease mutations and diminished by a surprisingly promiscuous inhibitor of the critical monomeric β-conformational change, neurotoxicity, and neurodegeneration. Hence, we postulate that specific mechanostable conformers are the cause of these diseases, representing important new early-diagnostic and therapeutic targets. The demonstrated ability to inhibit the conformational heterogeneity of these proteins by a single pharmacological agent reveals common features in the monomer and suggests a common pathway to diagnose, prevent, halt, or reverse multiple neurodegenerative diseases.

3. Hervás R, Galera-Prat A, **Gómez-Sicilia À**, Losada-Urzáiz F, Fernández MC, Fernández-Bravo D, Santana E, Barrio-García C, Melero C & Carrión-Vázquez M. *Nanomechanics of proteins, both folded and disordered.* In *Single-molecule studies of proteins*, pages 1–47. Springer, 2013.

Chapter 1
Nanomechanics of Proteins, Both Folded and Disordered

Rubén Hervás, Albert Galera-Prat, Àngel Gómez-Sicilia, Fernando Losada-Urzáiz, María del Carmen Fernández, Débora Fernández-Bravo, Elena Santana, Clara Barrio-García, Carolina Melero, and Mariano Carrión-Vázquez

4. Galera-Prat A, Hermans R, Hervás R, **Gómez-Sicilia À** & Carrión-Vázquez M. *Single-molecule force spectroscopy.* In *Atomic force microscopy in liquid: biological applications*, pages 157–187. Wiley, 2012.

6
Single-Molecule Force Spectroscopy

Albert Galera-Prat, Rodolfo Hermans, Rubén Hervás, Àngel Gómez-Sicilia, and Mariano Carrión-Vázquez

6.1
Introduction

This chapter aims to introduce the main concepts of single-molecule force spectroscopy (SMFS) by atomic force microscopy (AFM), essentially focusing on the basic methodology and its most relevant biological applications. Special emphasis will be paid to the criteria used to identify true single-molecule events (the basis of obtaining meaningful data and interpreting it correctly), as well as the main achievements of the technique. More specialized information with details of the experiments can be found in other reviews [1–4].

Since AFM-based SMFS has been so far mainly applied to proteins, this chapter focuses mostly on protein mechanics, although a short account on its application to other biopolymers can be found in Section 6.4.

The field of protein nanomechanics has made progress in three different fronts: intramolecular interactions (protein folding and unfolding), intermolecular interactions (protein–biomolecule), and membrane protein extraction (where intramolecular and intermolecular interactions occur simultaneously). This chapter is mainly focused in intramolecular studies, for which unambiguous single-molecule markers have been developed. In the case of intermolecular interactions, internal single-event markers are not yet available, while in mechanical extraction of membrane proteins, their unfolding and membrane unbinding cannot be easily decoupled.

5. Chwastyk M, Galera-Prat A, Sikora M, **Gómez-Sicilia À**, Carrión-Vázquez M & Cieplak M. *Theoretical tests of the mechanical protection strategy in protein nanomechanics.* Proteins: Structure, Function, and Bioinformatics 2014. 82(5):717–726.

Theoretical tests of the mechanical protection strategy in protein nanomechanics

Mateusz Chwastyk,[1] Albert Galera–Prat,[2] Mateusz Sikora,[1,3] Àngel Gómez-Sicilia,[2] Mariano Carrión-Vázquez,[2] and Marek Cieplak[1]*

[1] Laboratory of Biological Physics, Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, 02-668 Warsaw, Poland

[2] Instituto Cajal, Consejo Superior de Investigaciones Cientificas (CSIC), IMDEA Nanociencias and CIBERNED, Av. Doctor Arce, 37, 28002 Madrid, Spain

[3] Institute of Science and Technology Austria, Klosterneuburg, Austria

ABSTRACT

We provide theoretical tests of a novel experimental technique to determine mechanostability of proteins based on stretching a mechanically protected protein by single-molecule force spectroscopy. This technique involves stretching a homogeneous or heterogeneous chain of reference proteins (single-molecule markers) in which one of them acts as host to the guest protein under study. The guest protein is grafted into the host through genetic engineering. It is expected that unraveling of the host precedes the unraveling of the guest removing ambiguities in the reading of the force-extension patterns of the guest protein. We study examples of such systems within a coarse-grained structure-based model. We consider systems with various ratios of mechanostability for the host and guest molecules and compare them to experimental results involving cohesin I as the guest molecule. For a comparison, we also study the force-displacement patterns in proteins that are linked in a serial fashion. We find that the mechanostability of the guest is similar to that of the isolated or serially linked protein. We also demonstrate that the ideal configuration of this strategy would be one in which the host is much more mechanostable than the single-molecule markers. We finally show that it is troublesome to use the highly stable cystine knot proteins as a host to graft a guest in stretching studies because this would involve a cleaving procedure.

6. **Gómez-Sicilia À**, Galera-Prat A & Carrión-Vázquez M. *Folding landscape unveiled by force clamp spectroscopy*. In *Single molecule methods for studying protein folding*. Wiley, in press.

7. **Gómez-Sicilia À**, Sikora M, Cieplak M & Carrión-Vázquez M. *The universe of polyglutamine structures*. PLoS Computational Biology. Under review.

8. Wołek K, **Gómez-Sicilia À** & Cieplak M. *Determination of contact maps in proteins: a combination of structural and chemical approaches*. Journal of Chemical Physics. Submitted.

9. **Gómez-Sicilia À**, Wojciechowski M, Cieplak M & Carrión-Vázquez M. *Polyglutamine degradation by biological unfoldases*. In preparation.